ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis MSc Economics & Business

What Determines Change in Individual Football Skills?

Performance Prediction in Professional Football Using Machine Learning

Abstract

This study identifies the determinants of future performance levels of male professional football players by applying a variety of machine learning techniques. For this purpose, data from the football video game *FIFA* from 2007 until 2020 is used. With this data, a comparison between an Artificial Neural Network, Random Forest, and LASSO-regularised Linear Regression is performed. The results show that it is possible to predict the future performance level of football players fairly accurately. An Artificial Neural Network yields the best results when predicting future performance in one year, whereas a Random Forest gives the most accurate predictions when predicting future performance in three years. The results also show that a small number of football skills, such as standing tackle, sliding tackle, finishing, and marking, has the most profound impact on future performance. Which football skills are the most influential change per prediction period.

Keywords: machine learning, football, performance prediction, Artificial Neural Network, Random Forest, Bayesian Hyperparameter Optimisation

Student: Rogier de Bruin
Student ID: 451943

Supervisor: Dr. J.E.M. van Nierop
Second assessor: Dr. F. Frasincar

Date version: 31-07-2020

“We are card counters at the blackjack table. And we’re gonna turn the odds on the casino. (...) If we pull this off, we change the game. We change the game for good.”

-

Billy Beane

General Manager of the 2002 Oakland Athletics

Acknowledgements

I would like to thank my supervisor Dr. Van Nierop, whose efforts pushed this thesis to a higher level. This thesis was written during challenging and unprecedented times. At the time of the first meeting, no one would have expected that we were only months away from a complete lockdown caused by a global pandemic. From that point onwards, things changed drastically: university closed down, and digital communication became the only option. However, in my opinion, this has never felt as a constraint. The quality of the feedback was from an exceptional quality and extremely helpful. Besides the excellent quality of the feedback, the quantity was always right as well. It was just enough to bring me back on track, but not enough to let me incorporate certain ideas without overthinking it myself. Without his guidance and persistent help, the writing of this thesis would not have been possible.

Furthermore, I would like to thank all family and friends that have supported me in one way or another during this thesis process. Without any obligation to do so, I could always reach out for any sort of unconditional support. This has been a great help while writing this thesis and has definitely enhanced its quality.

Table of Contents

1. Introduction

With an estimated number of 3.5 billion fans (Das, 2020), football can be considered among the most popular sports worldwide. 380 million people have watched the 2018 Champions League final (Goble, 2019), and the revenue of multiple football clubs has exceeded 500 million euros in the 2018/2019 season (Lange, 2019). For the same season, the total European football market revenue equalled €28.9 billion (Deloitte, 2020). Stadiums are filled with people who pay high prices to support their club and who do not want to be disappointed by a team's performance. Coaches and technical directors have to ensure that a team performs at its best. Despite all this interest and effort, a major problem in the global sports industry is that information about the future performance of a player is lacking. An example will explain this. Imagine a person seeing a shiny and good-looking red apple in the supermarket, and he decides to buy the apple for breakfast the following day. As the person wakes up the following morning, the apple is not that nice and shiny anymore but has some bruises and bad spots on it. The person is disappointed by this, as he would not have bought the apple knowing this beforehand. The same holds for football players; there is no certainty about how well a player will perform in the future.

The story of the 2002 Oakland Athletics shows that statistical applications can change an entire sport (Lewis, 2004). Playing in the Major League Baseball (MLB), the Oakland Athletics had the third-lowest budget and, thus, had to compete against teams with more financial possibilities. By contracting players based on statistics, such as the on-base percentage (i.e., the ratio of having achieved a base to total at-bats), instead of subjective measures, which was common at that time, the Oakland Athletics achieved outstanding results. The team won the American League West title, only to be defeated in the American League Division Series (ALDS). To this day, they are the only team to have won 20 consecutive games in the MLB. The success of the Oakland Athletics shows that recognising undervalued players and exploiting this knowledge can make a team compete with financially superior opponents (Hakes & Sauer, 2006). Despite the success in baseball, Gerrard (2007) states that this method is less likely to succeed in complex invasion team sports, such as football. This is said to be due to problems that arise when trying to separate highly interdependent team plays into individual player actions. It is difficult to measure whether a striker scored a goal because of his excellent finishing skills, or because his

teammates play exceptionally well and deliver a good assist. The same question arises regarding opponents: the goalkeeper of the opposing team may perform poorly, allowing the striker to score relatively easily. This study tries to overcome the interdependency problem by using highly-detailed individual statistics of players.

The high level of uncertainty regarding the future performance of players causes significant problems for many. Three parties benefit immediately from a reduction in uncertainty. Firstly, football clubs profit from lower levels of uncertainty. Commonly, players are offered contracts for multiple years, despite it being unknown how well a player will perform in the future. For example, the development of a player that excels in youth teams may stop at a certain age, causing the player not to breakthrough at a professional level. The age at which the skills of a player start declining differs per player. These are just two examples of the uncertainty that football clubs face when contracting players. Football clubs worldwide can benefit from understanding how football players will develop in the future. Secondly, coaches benefit from having a better understanding of what aspects of football influence future performance of players. In this manner, they can adapt their training program to the aspects of football that positively influence future performance. For instance, at a younger age, more time can be spent to the aspects that are found to result in higher levels of overall quality of players in the future. More profound knowledge about what drives football talent will help players train more efficiently and, in the end, become a better player. Lastly, the uncertainty that many players face could be solved by a predictive model. Players can benefit from knowing what their future level will be, for example during salary negotiations. Also, youth players can have more insight into their future level and, based on this, determine whether they want to continue playing professional football. For instance, a player that is predicted to be not good enough for professional clubs at the age of sixteen can decide to focus on his educational career, rather than pursuing a football career. In the end, it can be stated that football clubs, coaches, and players benefit from experiencing lower levels of uncertainty about future performance of players.

Many articles have been written on performance prediction and player development. These studies are not necessarily focussed on sports but, for example, on professional expertise (Björkman, Ehrnrooth, Mäkelä, Smale, & Sumelius, 2013), musical capabilities (Baum, Owen, & Oreck, 1996), or the ability to learn something new (Feldhusen, 1994). Others do

specifically try to identify talents in sports. Mills, Butt, Maynard, and Harwood (2014) try to identify the effect of the environment that is created by football coaches on the development of youth players. Abbott and Collins (2004) state that the quality of the process of a player's development is the best predictor of talent and, therefore, advocate that a player's skills are continuously monitored. An overview of the main findings regarding future performance prediction is provided by Williams and Reilly (2000). Based on the number of articles, many people from all sorts of industries seemingly want to predict future individual skills. Despite the dissimilarities between the previously mentioned articles, they all seem to have one thing in common: future skills predictions are based on sociological, physiological, or psychological factors. Williams and Reilly (2000) are one of the few that mention the existence of a relationship between physical predictors and talent. In this case, however, physical predictors are limited to body type measures, such as length, weight, and body fat. An exception to this is the research by Reilly, Williams, Nevill, and Franks (2000), which states that 15 or 16 years old elite football athletes score higher on agility, sprint time, ego orientation, and anticipation skills than equally old sub-elite athletes. However, these conclusions should be handled cautiously, as only 31 athletes are used to investigate this difference. This illustrates the lack of scientific research into the relationship between sport-specific skills and the development of players, let alone that advanced machine learning techniques have been applied to this subject.

A significant problem found in many articles is that it is troublesome to gather detailed player information on a large scale. To exemplify this, the research by Reilly et al. (2000) required 28 different tests to be carried out onto 31 participants. Gathering detailed data is a time-consuming task, making it almost impossible to replicate the same research on a larger scale. It seems that the unavailability of data is one of the main reasons why this subject has not been researched more thoroughly. Reilly et al. (2000) mention privacy legislation and difficulties regarding the measuring of some subjective skills (e.g., mentality) as other reasons that limit the research possibilities in this field. For this study, data from the popular football video game *FIFA* will be gathered. This game contains detailed information about many professional football players around the world. As a result, plenty of data will be available to investigate the relationship between sport-specific skills and overall development of a player.

Since the sociological, physiological, and psychological factors of skills development have been investigated relatively often, the scope of this article will be limited to sport-specific factors. Also, due to the unavailability of data about women and non-professional players, only professional male football players will be taken into account. Therefore, the research question that will be investigated is:

> *What sport-specific player characteristics influence the future performance level of male professional football players?*

Sport-specific player characteristics are the abilities that a football player needs. Examples of this are how well a player heads, how fast he can run, or how well he defends. These characteristics are typically related to a player's performance level. This indicates at what level a player commonly performs. Some subquestions will help to answer the research question:

1) **How are player characteristics related to future performance?** To answer the research question, it is not only essential to predict a football player's future performance level accurately, but also to understand which factors drive the prediction. Both the direction and the size of the effect have to be taken into consideration.

2) **How does the variable importance differ when the period between the observation of a player's current level and characteristics of the past change?** Different models can be created by changing the future moment in time that has to be predicted. A model that predicts the level of a player in three years is probably less accurate than a model that predicts the progress in only one year. On the other hand, being able to predict over an extended period may give other relevant insights into the development of players. For example, football clubs may be interested in a longer-term prediction as well, despite the probably lower accuracy. It is possible that the importance of variables changes according to the different models.

3) **How does the relationship between variables and future performance level differ depending on the age of players?** It may be the case that different characteristics are responsible for the improvement and the decline of players of different ages. For example, the characteristics that predict the improvement of a young

player may not be the same as the ones that predict the decline of an older player. Therefore, it would be interesting to see if any changes occur based on the age of a player.

4) **How does the relationship between variables and future performance level differ depending on the preferred position of players?** Defenders likely require other skills than attackers do. Therefore, it is interesting to investigate whether the variable importance changes depending on the position of a player.

5) **Which machine learning model yields the highest predictive power?** Eventually, football clubs are most likely to be interested in the model that predicts player development most accurately. It has to be determined what kind of model achieves the highest accuracy.

Firstly, the existing literature regarding this subject will be discussed in the second section. This includes reviews of determinants of performance in football and machine learning in sports. After this, the data and methods used for this research are presented in the third and fourth section, respectively. In the fifth section, the obtained results will be provided. Lastly, the implications of the results will be presented in the sixth section, and a conclusion will be discussed in the seventh section.

2. Theory about performance prediction in sports

In this section, the existing literature about performance prediction in sports will be discussed. To start, an introduction to the determinants of performance in football will be provided. Factors that are negatively related to performance, as well as those that are positively related, will be discussed. This will be followed by an overview of existing articles that applied machine learning techniques in sports. Articles that use machine learning techniques to predict future performance will be reviewed in particular. Finally, a conceptual framework summarising the existing literature regarding future performance prediction will be presented.

2.1. Performance predictors in football

To understand what drives the future performance of football players, it is essential to gain insights into what determines their performance in general. According to Ali (2011), performance in football is determined by cognitive, perceptual and motor skills. Since football is a free-flowing game, many skills have to be executed in a dynamic context. That means that players not only have to perform their technique well but also at the right time. Therefore, football players should possess all three skills to perform at a high level. Besides this, it is also recognised that the skills of players decline from a certain age (Ali, 2011). Firstly, the positive relationship between football performance and cognitive, perceptual, and motor skills will be discussed. Additionally, the decline of player skills caused by both physical and mental causes will be reviewed. Lastly, an overview of the existing literature about the relationship between youth and future sports results will be provided.

2.1.1. Cognitive skills

Cognitive skills refer to how well a player understands football from a mental perspective (Ali, 2011). Cognitive skills are vital for football players, as they have to quickly assess all the aspects of the game that change around them (Williams & Reilly, 2000). Ali (2011) states that a challenge in measuring cognitive skills is identifying whether a player does not know, or does not recognise what to do. It is a sign of lacking cognitive skills if a player does not know what to do. Nevertheless, it may be possible that a player knows which actions to perform in certain situations, but is unable to recognise this. In this case, the player's cognitive skills are well-developed, but his perceptual skills, which will be examined in the following section, are lacking. Generally, it can be stated that the cognitive skills of a player are more developed when a player is more experienced (Ali, 2011; Elferink-Gemser, Visscher, Richart, & Lemmink, 2004).

2.1.2. Perceptual skills

Perceptual skills indicate how well players can react to specific actions and movements in their surroundings (Ali, 2011). This is commonly measured by showing players a variety of football situations and ask them what they would do, after which their answers are compared to those of an experienced football coach. A problem regarding measuring perceptual skills in this manner is that it will always be subjective what the right decision is (Ali, 2011). There may be a disagreement between coaches what the right decision is in a

certain situation. Despite this, multiple studies have found that perceptual skills are positively related to player experience, as is the case for cognitive skills (McMorris & Graydon, 1996; McMorris & Graydon, 1997; Williams & Davids, 1998). This is confirmed by a model predicting the adult performance level of football players based on positioning and deciding skills at the age of 17 achieving an accuracy of 70% (Kannekens, Elferink-Gemser, & Visscher, 2011).

2.1.3. Motor skills

Motor skills describe how well a player can control his body movements. As such, motor skills are directly related to, amongst others, passing, controlling, dribbling and shooting the ball. Research has shown that a positive relationship between motor skill competence and general fitness exists (Haga, 2009). It is also proven that this relationship can be used to predict future general fitness levels; children with well-developed motor skills are more likely to have a better general fitness at a later age (Barnett, van Beurden, Morgan, Brooks, & Beard, 2008). Research by Di Cagno et al. (2014) demonstrates this as well by showing that gymnasts of approximately twelve years old with well-developed motor skills achieve better results three years later than gymnasts with less-developed motor skills. Furthermore, 79% of the variance in general fitness can be explained by a person's ability to jump, throw, and kick (Stodden, Langendorfer, & Roberton, 2009). Especially jumping can be considered an essential predictor for fitness, while kicking seems less critical. Although general fitness levels do not cover all aspects of football, players do need the ability to perform many physically challenging movements during the game. Besides a positive relationship between motor skills and general fitness, it is found that top-level players are better in juggling the ball than other players (Hoare & Warr, 2000; Rösch, et al., 2000). Also, elite football players appear to have a better balance ability in comparison to less proficient players (Hrysomallis, 2011). Furthermore, more advanced motor skills lead to a better kicking technique (Stratton, Reilly, Williams, & Richardson, 2004). Additionally, Gonaus and Müller (2012) state that youth football players who perform better in speed, power and flexibility, and coordination and endurance are more likely to be drafted later in their career. In the end, it appears that there is a positive relationship between motor skills and football skills. An important notion is that the required motor skills differ per position (Di Salvo, et al., 2007). For example, it is shown that midfielders need more running capabilities than players in other positions.

The exact determinants of a player's motor skills are not so clear, however. Many studies show that most motor skills are primarily developed before the age of six (Chow, Henderson, & Barnett, 2001; Iivonen, Sääkslahti, & Nissinen, 2011; McKenzie, et al., 2002). In this short period that motor skills improve most, habitual and frequent physical activity is a significant predictor (Bürgi, et al., 2011; Cliff, Okely, Smith, & McKeen, 2009; Fisher, et al., 2005; Williams, et al., 2008). This means that the extent to which children gain control over their body movements during early childhood is a determinant for motor skills later at a later age. Furthermore, a negative relationship exists between a person's motor skills and certain medical conditions during early childhood, such as prematurity, intrauterine growth restriction, many hospitalisations, and immobility (Sanders-Woudstra, Verhulst, & De Witte, 1993). In the end, it can be stated that multiple circumstances during early childhood affect a person's motor skills during the remainder of his life. As such, a significant part of the performance of football players is determined by childhood development.

2.1.4. Physical strength

In this research, the causes of the decline of individual skills are divided into two parts: physical and mental causes. Starting with physical causes, an important aspect regarding the decline of individual skills is a player's peak performance age. This is the age at which an athlete performs best. Before the peak performance age, the performance of an athlete is still improving, while the performance is deteriorating after the peak performance age. The peak performance age differs per sport and athlete, making it hard to accurately state the age at which an athlete excels (Baltes & Baltes, 1990). In general, athletes that participate in sports that require a high level of explosiveness peak at a relatively young age, while athletes that are participating in endurance sports peak later (Schulz & Curnow, 1988). The same research states that, since football can be considered a sport that requires explosiveness, the peak performance age of a football player is around 25 years. The main reason why the performance of older players deteriorates is the ageing of skeletal muscles, which causes the numbers of fibres in muscles starting to diminish (Faulkner, Davis, Mendias, & Brooks, 2008). This has an immediate effect on the motor skills of football players.

2.1.5. Mental strength

Secondly, there are mental causes that can lead to a decline in a player's skills. There is a broad variety of articles that supply different ages at which cognitive decline starts, ranging

from 18 to 70 years old. Some papers suggest that mental decline starts at a relatively high age. Examples of this are that decline begins when people are 70 years old (Aartsen, Smits, Van Tilburg, Knipscheer, & Deeg, 2002), 60 years old (Plassman, et al., 1995), or 55 years old (Rönnlund, Nyberg, Bäckman, & Nilsson, 2005). These conclusions are based on multiple cognitive tests per person. However, Salthouse (2009) suggests that most of these findings are unreliable due to the longitudinal comparisons that are made. As a result of this, the age-related decline is said to be masked by positive effects associated with prior test experience. According to Schroeder and Salthouse (2004), cognitive decline starts in a person's early 20's. Many others support this age (Allen, Bruss, Brown, & Damasio, 2005; Fotenos, Snyder, Girton, Morris, & Buckner, 2005; Salat, et al., 2004; Sowell, et al., 2003; Sullivan & Pfefferbaum, 2006). Therefore, it seems more likely that cognitive skills start to decline slowly at the age of 20. The decline accelerates when a person reaches the age of approximately 50 years old. Although professional football players are almost always younger than 50, this implies that their cognitive and perceptual skills could gradually decline over the years. However, this appears paradoxical to the earlier mentioned finding that cognitive and perceptual skills improve when players become more experienced. It is plausible that both statements are correct, but that the effect of being more experienced is stronger than the effect of the gradual mental decline.

It must be noted that the two causes of individual decline seem certain. The preceding findings leave little to no doubt that a player will deteriorate both physically and mentally starting at a particular age. It is impossible to stop the process of an ageing body. Although the deterioration age differs per player and is unknown on beforehand, the performance of all players will certainly decline in the future. However, it is not certain that a player's performance will improve at any point in time. Despite cognitive and perceptual skills being related to experience, it is not certain that a player's performance will increase. As a result, it can be argued that predicting the decline of performance is easier than predicting the increase. A decline will certainly occur at some point in time, whereas it is unclear if a player will better its performance. As such, it can be stated that predicting decline is not about whether a player's performance will decrease, but when they will decrease and at what rate.

2.1.6. Youth results

Besides all other aspects that are related to future performance, it appears sensible that athletes who perform well at a young age, are more likely to obtain good results in the future as well. However, excelling in a specific sport as a child appears to be unrelated to performing well at a later age (Elferink-Gemser, Jordet, Coelho-E-Silva, & Visscher, 2011). This is confirmed by the findings that youth results in tennis are not a good predictor for success later on (Brouwers, de Bosscher, & Sotiriadou, 2012). A possible explanation for this may be the relative maturity of youth players. It is shown that attributes in which youth players can have only a temporary advantage over other players are not related to career success (Unnithan, White, Georgiou, Iga, & Drust, 2012). Examples of such temporary advantages are higher body mass and bigger stature at a young age. Unnithan et al. (2012) state that these attributes are related to kicking with more force and a higher vertical jump capacity. A youth player will be better than other youth players from the same age in these two skills due to its relatively early physical maturity. In the long term, however, this advantage will disappear since his peers will catch up to him. As such, it can be stated that achieving good results at a certain age is not necessarily a good predictor for future success. Nevertheless, the background of why a player achieves these results could be a determinant.

2.2. Machine learning in sports

In the preceding sections, the determinants of future performance have been discussed. Before machine learning can be applied to predict future performance of football players, it is vital to explore how other studies have tried to use machine learning in sports. Although statistics have played a role in sports for a long time, the use of advanced machine learning techniques is less common. Machine learning techniques use computer algorithms that improve through experience. It can be considered a subset of artificial intelligence. According to Bunker and Thabtah (2019), Purucker (1996) was one of the first to use machine learning techniques predicting results in the National Football League (NFL) based on five team features. The Artificial Neural Network (ANN) constructed by Purucker achieved an accuracy of 61%, which is lower than the 72% accuracy achieved by experts. An extended version of Purucker's model, however, managed to perform better than NFL experts (Kahn, 2003). The ANN achieved an accuracy of 75%, compared to an accuracy of 63% by experts. Although the results are relatively accurate, both studies use a relatively small data set of 64 and 208 matches, respectively. A more extensive study gathered data

about all matches in four major league sports (i.e., National Rugby League (NRL), Australian Football League (AFL), Super Rugby and English Premier League (EPL)) starting from 2002 (McCabe & Trevethan, 2008). The constructed ANN achieved an accuracy of approximately 67.5%. According to McCabe and Trevethan (2008), this is higher than the accuracy of sports analysts' predictions, who achieved an accuracy between 60% and 65%. Another study uses Bayesian inference and rule-based reasoning to predict the outcome of the 2002 World Cup Soccer (Min, Kim, Choe, Eom, & McKay, 2008). Although it is hard to calculate exact numbers about the achieved accuracy due to the nature of their study, the predictions appear reasonable and stable. It must be noted, however, that advanced machine learning techniques do not necessarily outperform less complicated models in these types of study. A comparison between a Logistic Regression, ANN, Support Vector Machine (SVM) and Naive Bayes shows that the Logistic Regression performs best when predicting National Basketball Association (NBA) outcomes (Cao, 2012).

As mentioned before, the 2002 Oakland Athletics achieved outstanding results in the MLB by using machine learning techniques. The main idea was that teams with lower budgets could compete with the high budget teams by buying undervalued players (Lewis, 2004). It would be expected that a player's salary reflects his performance. However, one of the main findings was that, at that time, on-base percentage (i.e., the ratio of having achieved a base to total at-bats) is a significant predictor of a team winning, but not of player salaries. Traditionally, teams often considered a player's batting average (i.e., the ratio of hits to total at-bats) or slugging rate (i.e., the ratio of bases reached to total at-bats). The main difference is that the on-base percentage also accounts for walk-offs, while both batting average and slugging rate disregard this. As a result, players with a high on-base percentage were often undervalued. Similarly, it was found that the same holds for closers (i.e., a pitcher being specialised in getting the final outs in a close game when his team is leading) being overrated and overpaid. Although many baseball experts doubted these statements, they are confirmed by scientific research (Hakes & Sauer, 2006). It must be noted, however, that these abnormalities in player's salaries were quickly resolved. It is shown that on-base percentage was a better predictor for salary than slugging rate in 2004 (Hakes & Sauer, 2007). This indicates that it is vital to update predictive models continuously and to stay ahead of the curve in order to identify undervalued players.

The above mentioned studies all applied machine learning on a team level. The performance of an athlete can be predicted by machine learning techniques on an individual level as well. The performance of a javelin thrower has been predicted by both a regression model as well as an ANN (Maszczyk, et al., 2014). It is found that the ANN yields more accurate results than the regression model. Additionally, the result of the 200-meter backstroke women's final at the Olympic Games in 2000 has been predicted by using an ANN (Edelmann-Nusser, Hohmann, & Henneberg, 2002). The error of the prediction was only 0.05 seconds on a total swimming time of more than two minutes. Both studies underline the importance of their results for high-performance staff to identify the factors that determine success. In biathlon competitions, it is found that predicting shooting hit rates using a Tree-based model with Boosting is more successful than using a Logistic Regression or ANN (Maier, Meister, Trösch, & Wehrlin, 2018).

In sum, it appears that ANNs in particular can predict performance in sports relatively accurately, both for individuals as for teams. Moreover, ANNs are often found to be more accurate than other models. However, some studies indicate that other methods are more reliable than ANNs. Therefore, it cannot be assumed that an ANN is necessarily the best method to predict performance in sports, indicating that other methods have to be considered as well.

2.3. Conceptual model

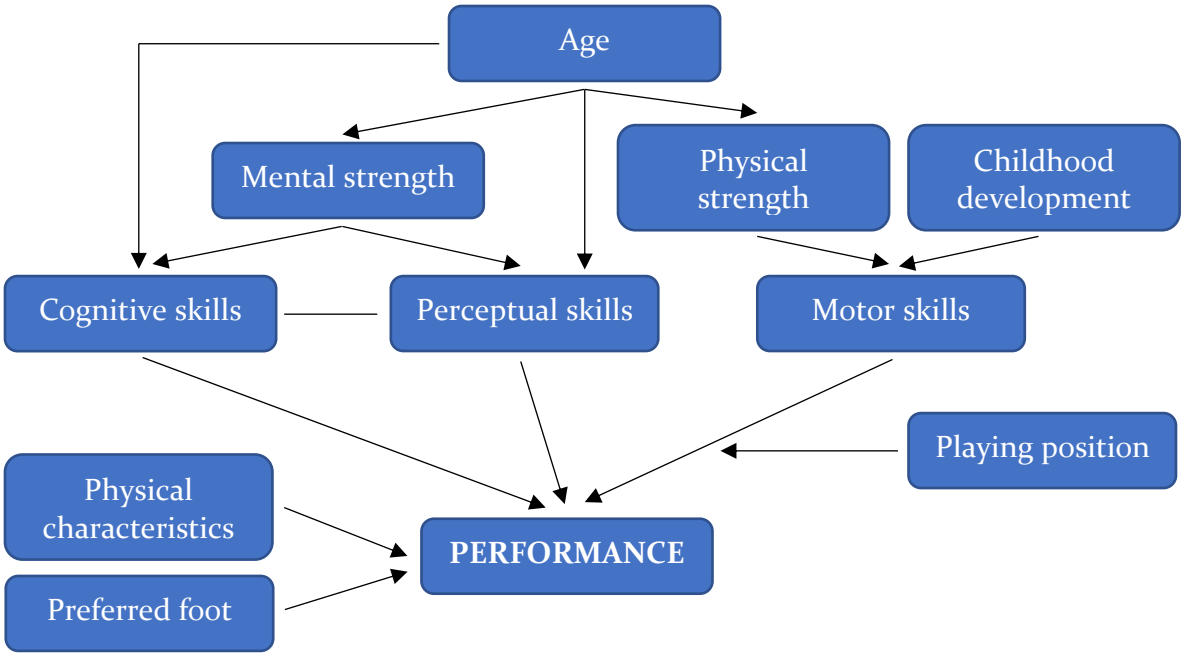


Figure 1: Conceptual model

To summarize the existing literature about performance prediction in sports, a conceptual model is presented in **Error! Reference source not found.**. As stated in Section 2.1, cognitive, perceptual, and motor skills are the main predictors of performance in football (Ali, 2011). Cognitive (Ali, 2011; Elferink-Gemser, Visscher, Richart, & Lemmink, 2004) and perceptual skills (Kannekens, Elferink-Gemser, & Visscher, 2011; McMorris & Graydon, 1996; McMorris & Graydon, 1997; Williams & Davids, 1998) are positively related to experience, which is highly correlated to age. This means that older football players are cognitively and perceptually more skilled. Mental strength is positively related to both cognitive and perceptual skills as well and depreciates as a player becomes older (Schroeder & Salthouse, 2004). Since age is positively related to cognitive and perceptual skills but negatively related to mental strength, it is unclear how cognitive and perceptual skills develop over the years. It is suggested, however, that the positive effect of experience is stronger than the negative impact of mental strength. As a result of cognitive and perceptual skills being related to the same factors, these two skills are highly inter-related. Players with well-developed cognitive skills are likely to have excellent perceptual skills as well, and vice versa.

Furthermore, motor skills are positively related to childhood development. Detrimental circumstances during childhood will result in lower motor skills level at a later age (Bürgi, et al., 2011; Cliff, Okely, Smith, & McKeen, 2009; Fisher, et al., 2005; Sanders-Woudstra, Verhulst, & De Witte, 1993; Williams, et al., 2008). Additionally, there is a positive relationship between physical strength and motor skills (Faulkner, Davis, Mendias, & Brooks, 2008). Physical strength, however, is negatively related to age (Schulz & Curnow, 1988). As soon as a player has passed its peak performance age, his skeletal muscles will deteriorate (Baltes & Baltes, 1990). This will have an immediate effect on his motor skills, indicating that these will decline when a player becomes older.

Lastly, it is recognised that a player's position may influence the effect of these skills on performance (Di Salvo, et al., 2007). For example, defenders are likely to require a different set of skills than goalkeepers, midfielders and attackers. Also, it is necessary to control for physical characteristics and which foot a player prefers when measuring performance. Especially for young players, the maturity rate differs per player (Baltes & Baltes, 1990). According to the existing literature, this may affect the performance development for players. Although there is no literature found that suggests a relationship between the

preferred foot of a player and its performance, it is plausible that this relationship exists. Therefore, it is accounted for when constructing the model.

3. Data

To investigate the research question, a data set containing detailed information about football players is constructed. The information is gathered from multiple editions of the popular football video game *FIFA*, in which gamers can control and play with real-life football players. Only professional male football players that are active in the major football leagues are included in the game. Since it differs per year which leagues are covered, an overview of the leagues per year is depicted in Table 6 (Appendix A). In addition to all players active in these leagues, players from approximately ten clubs from other leagues are included. Again, it varies per year which teams are chosen. The data in *FIFA* is constructed by approximately 6,000 FIFA Data Reviewers, who monitor the development of individual football players. As a result, the data is highly accurate and closely related to the performance of a player in reality. A player's capabilities are based on a variety of variables, such as heading accuracy, agility, and aggression. Player attributes are stored in *FIFA* as well, which include, for example, a player's age and physical characteristics.

Data about all football players present in *FIFA* from 2007 until 2020 is merged in a data set with 158,062 observations, featuring 39,719 players. These players have 179 different nationalities from all continents except Antarctica. It is important to note that there is a difference between the calendar year and the release date of a *FIFA* edition. For example, *FIFA 20* is released on the 24th of September 2019. Although the information in the game is updated continuously, the gathered data is always from the release date of the game. The data can be divided into player attributes and football data. To start with the former, the available player attributes are *age*, *height*, *weight*, *preferred foot* and *preferred position*. *Age* is measured in years, *height* in centimetres, and *weight* in kilograms. *Preferred foot* indicates whether a player is right-footed or left-footed. Lastly, *preferred position* indicates in which position a player is most often active. Secondly, football data is stored in the data set. This consists of both *attacking* and *defensive work rate*, an *overall* score, and 33 football-specific skills, including goalkeeper-related skills. *Attacking* and *defensive work rate* are categorised into 'low', 'medium', or 'high'. All other football data is scaled from 1 to 99 with 1 being the

lowest score. Descriptive statistics of the continuous and categorical variables are shown in Table 7 and Table 8 (Appendix B), respectively.

In the previous section, a conceptual model of football performance was presented. Variables representing most concepts of the model are present in the data set. However, some variables in the model cannot be found in the data. This is visualised in

Figure 2. The green boxes in the model represent variables that can be observed in the data set, whereas the red boxes denote unobserved variables. The main reason for those variables not being present in the data set is the difficulty to measure them. These variables are not possible to review during matches without some sort of interactions with a player. Therefore, these variables are not included in this study.

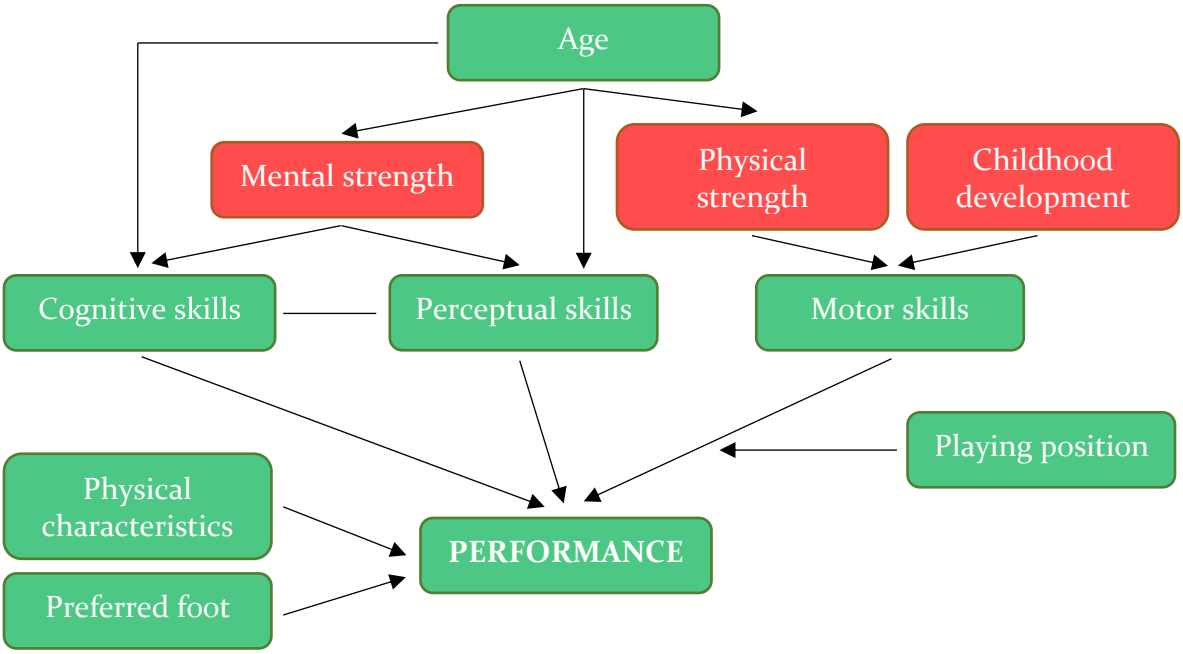


Figure 2: Overview of observed (green) and unobserved (red) variables in the conceptual model

Most variables contained some missing values. This was mainly due to FIFA Data Reviewers being unable to determine these values. 195 missing values belonged to players that did not play any matches. As their skills have not been assessed, these observations contained only missing values. Additionally, some skills have not been evaluated for certain players, as FIFA Data Reviewers did not have enough information about the skills of a player. This was mainly the case for some highly specific skills, such as goalkeeping skills and volleys. Lastly, *FIFA* did not always include a player's preferred position before 2015. As a result, the preferred position of players active before 2015 was missing relatively often. In total, 13,883

observations of 1,367 players contain missing values. The observations containing missing variables are kept in the data set momentarily. Eventually, after some other data processing steps, the observations with missing values are removed.

4. Methods

This study aims to predict future performance levels of football players and to find out what drives these predictions. The main method that is used to predict future performance levels is a feed-forward ANN. Section 2.2 showed that this method is most likely to result in the most accurate predictions. However, Section 2.2 also indicated that the results should be compared to other machine learning methods, as there is no certainty that this method is the best. Therefore, a Random Forest and LASSO-regularised Linear Regression are constructed to compare the performance of the ANNs. For this study, two different prediction periods are considered: a period of one year, and a period of three years. A separate model is created per prediction period. In the first part of this section, the theory of three separate machine learning methods will be explained. In the remaining parts, the application of the theory is discussed.

4.1. Theoretical basis of the models

In this section, the theory about the machine learning methods that are used will be discussed. Firstly, the theory behind an ANN will be provided. Afterwards, the same will be done for a Random Forest and a Regularised Linear Regression.

4.1.1. Artificial Neural Network

The idea of a Neural Network has been brought forward in an article by McCulloch and Pitts (1943). They made use of electrical circuits to develop a simple Neural Network. Over the decades, progress in research led to the development of more complicated Neural Networks. An example of this is the construction of an Artificial Neural Network (i.e., a Neural Network created with the help of computers) by Rochester, Holland, Haibt and Duda (1956). A schematic overview of the architecture of an ANN is visualised in Figure 3. An ANN aims to create a set of linkages between the input layer and the output layer of the model. The input layer consists of raw data, whereas the output layer contains the eventual prediction. The value in the output layer is determined by passing the input data through several hidden

layers. The number of hidden layers in an ANN creates a trade-off between accuracy and computational costs. Additional hidden layers may improve the prediction accuracy, but also enlarges the training time. Generally, two hidden layers are sufficient, but three hidden layers can be used when the main aim of the model is to achieve a high prediction accuracy (Karsoliya, 2012). Adding a fourth hidden layer is shown to increase computational costs without increasing prediction accuracy. Therefore, the number of hidden layers is held constant at three for this study, which is the same number of hidden layers as displayed in Figure 3. The circles that are visible in Figure 3 are referred to as neurons and determine the actual predicted value. Every hidden layer contains a pre-set number of neurons. This number is a hyperparameter, which means that it has to be set manually before the training process begins. Hyperparameters have to be optimised to achieve the best results.

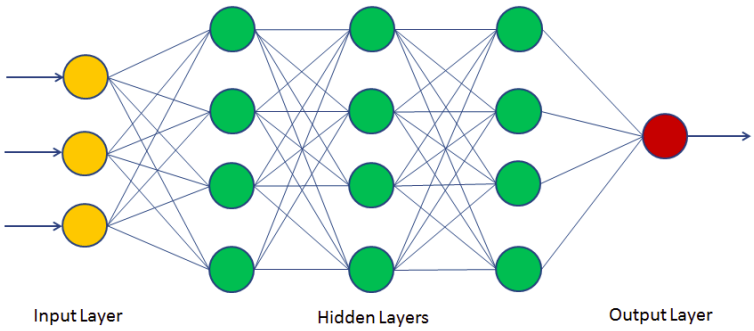


Figure 3: Schematic representation of the architecture of an Artificial Neural Network

The neurons in an ANN work similarly to the neurons in a human brain. All neurons are connected by dendrites, through which information is transmitted. The neurons receive information, transform it, and subsequently send it to the next neuron. This process is schematically depicted in Figure 4. In this figure, three inputs x_i are visualised. All inputs are assigned a weight w_i. This weight reflects the importance of the connection between two neurons. A higher weight assigns more importance to a particular input. The value of the neuron z is calculated as the weighted sum of the input, including a bias term b. This is mathematically depicted in equation (1):

$$\begin{aligned} \boldsymbol{z}_l &= \boldsymbol{w}_l^T \boldsymbol{a}_{l-1} + \boldsymbol{b}_l \\ \boldsymbol{a}_l &= f(\boldsymbol{z}_l) \end{aligned} \tag{1}$$

In this equation, $\boldsymbol{a}_l$ is the activation vector of hidden layer l and $\boldsymbol{z}_l$ the weighted input to neurons in hidden layer l. The weighted input is dependent on the transpose of the weights $\boldsymbol{w}_l^T$ of hidden layer l, the activation vector of the previous hidden layer $\boldsymbol{a}_{l-1}$, and the bias

term $\boldsymbol{b}_l$ of hidden layer l. The weighted sum is passed through the activation function f, which results in the eventual output.

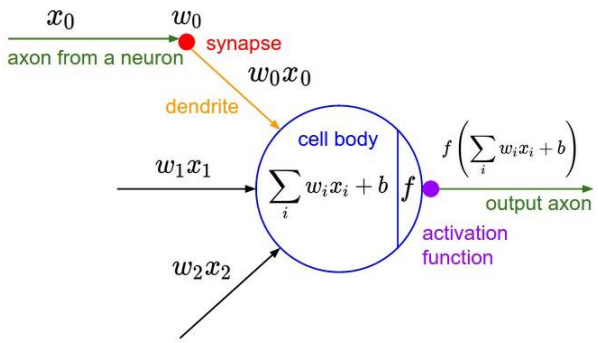


Figure 4: Schematic representation of the process involving a neuron

With only one hidden layer and one neuron, the mathematical steps are relatively simple. However, when multiple hidden layers and neurons are connected and transformed by non-linear functions, complex and non-linear relationships can be captured. At the same time, this increases the chance of overfitting. Overfitting means that a model performs well on predicting the training data, but poorly on unseen test data. This is a result of the model no longer being trained by the training data, but simply memorising it. To overcome this problem, dropout between two hidden layers can be applied. This indicates that a percentage of the neurons is randomly set to zero. The percentage, also referred to as dropout rate, is a hyperparameter.

A variety of activation functions exist to transform the weighted sums of neurons. For this study, the Rectified Linear Units (ReLU) activation function is used, as proposed by Nair and Hinton (2010), for both the hidden layers and the output layer. This activation function is chosen as it is proven to be a good default choice (Goodfellow, Bengio, & Courville, 2016). The ReLU activation function is depicted mathematically in equation (2):

$$f(z) = \max(0, z) \tag{2}$$

In this equation, z denotes the weighted input to a neuron.

In the end, the goal of the ANN is to determine the weights between the layers of the model. This is accomplished by minimising a loss function $\mathcal{L}$. Due to the size of an ANN, it is hard to define the optimum of $\mathcal{L}$ analytically. Therefore, a method called backpropagation is applied to find the optimum of the loss function, as proposed by Rumelhart, Hinton and

Williams (1985; 1986). This method propagates the error of the model backwards through the network. This means that data passes through the network forwards and backwards more than once. One full passage of all data is referred to as an epoch. When passing through the network, the data can be split into batches that pass through the network independently. The number of observations that each batch contains is the batch size. The number of epochs and batch size are hyperparameters. Backpropagation tries to identify the optimum of the loss function by computing its partial derivatives $\frac{\partial \mathcal{L}}{\partial w_{j,k,l}}$ and $\frac{\partial \mathcal{L}}{\partial b_{j,l}}$ with respect to any weight $w_{j,k,l}$ between neurons j and k in hidden layer l or bias $b_{j,l}$ of neuron j in hidden layer l. Subsequently, these partial derivatives are used to minimize the loss function $\mathcal{L}$. It is important to note that $\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_l} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_l}\frac{\partial \boldsymbol{a}_l}{\partial \boldsymbol{z}_l}\frac{\partial \boldsymbol{z}_l}{\partial \boldsymbol{W}_l}$ and $\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_l} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_l}\frac{\partial \boldsymbol{a}_l}{\partial \boldsymbol{z}_l}\frac{\partial \boldsymbol{z}_l}{\partial \boldsymbol{b}_l}$ both contain the component $\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_l}\frac{\partial \boldsymbol{a}_l}{\partial \boldsymbol{z}_l}$. As a result, the local error $\varepsilon_{j,l}$ of neuron j in hidden layer l can be defined as shown in equation (3):

$$\varepsilon_{j,l} \equiv \frac{\partial \mathcal{L}}{\partial \boldsymbol{z}_{j,l}} \tag{3}$$

Then, the error for each layer has to be calculated. This error can be propagated through the network backwards. This means that the error $\varepsilon_{j,N+1}$ for layer $\boldsymbol{l}^{N+1}$ has to be defined. This is visualised in equation (4):

$$\varepsilon_{j,N+1} = \frac{\partial \mathcal{L}}{\partial a_{j,N+1}} f'(z_{j,N+1}) \tag{4}$$

Equation (4) consists of two terms. The first term measures the degree of change in the loss function as a function of the j-th output activation. The second term represents the derivative of the activation function used in the output layer as evaluated in $z_{j,N+1}$. Equation (4) is rewritten in matrix-based form in equation (5) to ease further derivation of the backpropagation formulas:

$$\boldsymbol{\varepsilon}_{N+1} = \nabla_a \mathcal{L} \odot f'(\boldsymbol{z}_l) \tag{5}$$

In this equation, $\nabla_a \mathcal{L}$ denotes a k-dimensional vector consisting of the partial derivatives $\frac{\partial \mathcal{L}}{\partial a_{j,N+1}}$ for all k neurons in the output layer. Similarly, $f'(\mathbf{z}_l)$ represents the partial derivatives $f'(z_{j,N+1})$ for all j. The symbol $\odot$ denotes the Hadamard product of the terms on the left and the right. Based on equation (5), equation (6) represents the error ε_l in layer l as a function of the error ε_{l+1} in the next layer:

$$\varepsilon_l = ((\boldsymbol{W}_{l+1})^T \varepsilon_{l+1}) \odot f'(\boldsymbol{z}_l) \tag{6}$$

In this equation, $\boldsymbol{W}_{l+1}$ represents the weight matrix of the $(l+1)$-th layer, containing all weights between the l-th and the $(l+1)$-th layer. By combining equation (1), (5) and (6), the partial derivatives $\frac{\partial \mathcal{L}}{\partial w_{j,k,l}}$ and $\frac{\partial \mathcal{L}}{\partial b_{j,l}}$ can be represented as shown in equation (7):

$$\frac{\partial \mathcal{L}}{\partial w_{j,k,l}} = a_{k,l-1}\varepsilon_{j,l}$$
$$\frac{\partial \mathcal{L}}{\partial b_{j,l}} = \varepsilon_{j,l} \tag{7}$$

As the partial derivatives are now calculated through backpropagation, the loss function can be minimised using iterative algorithms, of which the Gradient Descent (GD) method is mostly used. The GD algorithm aims to optimise the parameters that are used by the network. This is schematically represented in Figure 5. It accomplishes this by calculating the loss for $\mathcal{L}(\theta)$, in which θ denotes a set of parameters and then take a small step of the size of the learning rate α towards the negative gradient. This is mathematically depicted in equation (8):

$$\theta_{new} = \theta_{old} - \alpha \nabla_{\theta} \mathcal{L}(\theta) \tag{8}$$

In this equation, $\nabla_{\theta}\mathcal{L}(\theta)$ denotes the gradient of $\mathcal{L}(\theta)$. A disadvantage of the GD algorithm is that, since this is an iterative algorithm, the computation time is high. Therefore, Stochastic Gradient Descent (SGD) is often applied to reduce computation time. SGD uses only a random sample of n' observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{n'}\} \in \boldsymbol{X}$ to calculate the gradient instead of all observations. Given that n' is sufficiently large and under the assumption that $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta)$, equation (9) holds:

$$\nabla_{\theta}\mathcal{L}(\theta) = \frac{\sum_X \nabla_{\theta,x}\mathcal{L}(\theta)}{n} \approx \frac{\sum_{j=1}^{m'} \nabla_{\theta,x_j}\mathcal{L}(\theta)}{m'} \tag{9}$$

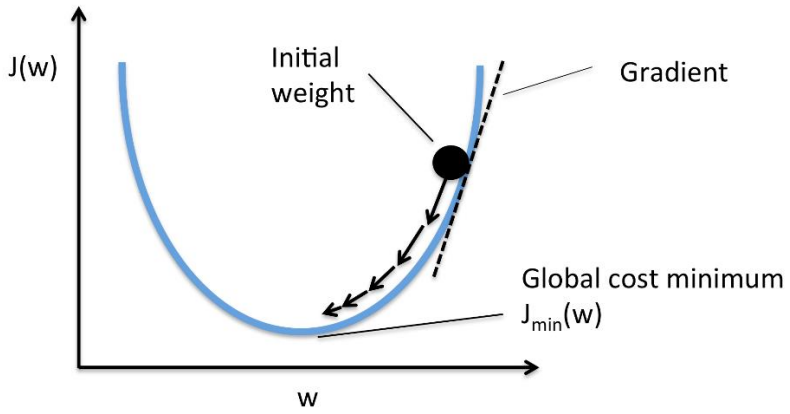


Figure 5: Schematic representation of Gradient Descent method

A disadvantage of (Stochastic) Gradient Descent, however, is that the learning rate α has to be chosen manually. Other optimisation techniques are capable of automatically adapting the learning rate and finding individual learning rates per parameter. Therefore, the Adaptive Moment (Adam) optimiser is used for this study, as proposed by Kingma and Ba (2014). The main idea behind the Adam optimiser is that it uses estimations of the first and second moments of the gradient to adapt the learning rate for every weight of a Neural Network. The N-th moment m_n of a random variable x is equal to the expected value of that variable to the power n. This is formalised in equation (10):

$$m_n = E[x^n] \tag{10}$$

To calculate the moments, the Adam optimiser uses the moving average of the gradient m_t and the squared gradient v_t. The calculation of the moving averages is depicted in equation (11):

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \end{aligned} \tag{11}$$

In this equation, β_1 and β_2 are hyperparameters, and g_t represents the gradient on the current random sample of observations. Experiments with multiple types of methods – including ANNs – have shown that values of 0.9 for β_1 and 0.999 for β_2 yield good results (Kingma & Ba, 2014). Therefore, β_1 and β_2 are held constant at 0.9 and 0.999, respectively.

The first iteration always uses vectors of zeroes as the moving averages. As a result, there is substantial bias in the moving average for small numbers of t and the calculated moving averages have to be corrected. This is depicted in equation (12):

$$\begin{aligned} \widehat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_\mathrm{t} &= \frac{v_\mathrm{t}}{1-\beta_2^\mathrm{t}} \end{aligned} \tag{12}$$

The weights of the model can now be updated using the moving averages. The weights are updated as shown in equation (13):

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta \frac{\widehat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{13}$$

In this equation, $\mathbf{w}_\mathrm{t}$ represents the model weights after iteration t, η is a hyperparameter and ε the error term. The same experiments that determined the values of β_1 and β_2 also indicate that a value of 0.001 for η yields good results for a variety of methods (Kingma & Ba, 2014). Therefore, the hyperparameter η is held constant at 0.001 for this research.

4.1.2. Random Forest

The Random Forest algorithm, as developed by Breiman (2001), is a tree-based ensemble method. It is based on the tree-based models of Breiman, Friedman, Stone and Olshen (1984) and extends the bagging algorithm of Breiman (1996). As the Random Forest algorithm is an ensemble method, multiple decision trees are constructed and the eventual predicted value is determined by combining the trees. An important distinction compared to other ensemble methods is that the Random Forest algorithm tries to reduce the correlation between the distinct decision trees. This is accomplished by using a randomly selected data set per decision tree, and a randomly selected number of variables for each split.

Decision trees aim to split the data set based on several conditions. All connected parts of a decision tree are referred to as nodes. The first node of a tree is named the root node, whereas the last nodes are called leaf nodes. Any nodes between the root node and the leaf nodes are referred to as internal nodes. All nodes are connected by branches. Although most tree-based algorithms allow a node to be split into more than two nodes, the Random Forest algorithm does not. A schematic representation of a decision tree is shown in Figure 6. The predicted value of an observation is the value of the leaf node that it belongs to. The value of the leaf nodes is the average of all observations in that node. At every split, the residual sum of squares (RSS) is minimised. The RSS is represented mathematically in equation (14):

$$RSS_j = \sum_{i=1}^{n_j} (y_j - y_i)^2 \tag{14}$$

In this equation, n_j refers to the number of observations in node j, y_j is the predicted value of observations in node j and y_i the actual value of observation i.

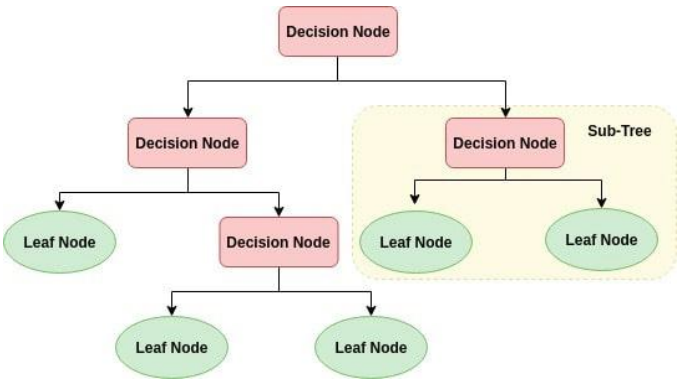


Figure 6: Schematic representation of a decision tree

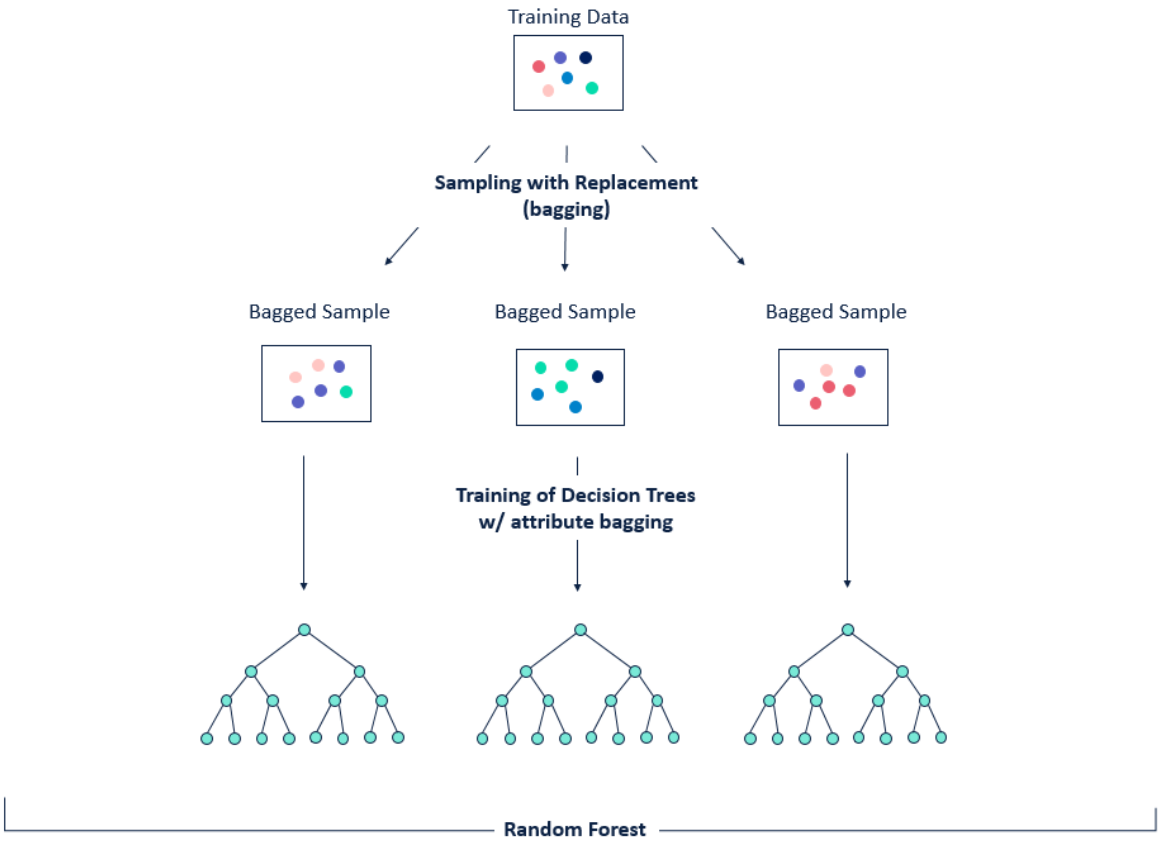


Figure 7: Schematic representation of the Random Forest algorithm

A schematic representation of the Random Forest algorithm is depicted in Figure 7. The algorithm uses two methods to reduce the correlation between the distinct trees. Firstly, a different randomly selected set of observations is used per tree. This set is selected through bootstrapping. This means that B bootstrap samples $\boldsymbol{X}^{*b}$ with the same size as the original data set are created. The bootstrap samples are filled with replacement, meaning that an observation can occur in a sample more than once, or not at all. As all decision trees are based on different data, there is less correlation between the trees. For all bootstrap samples, the predicted value $\hat{f}^{*b}(\boldsymbol{X})$ is calculated, in which f^{*b} denotes the decision tree of bootstrap sample b. The eventual predicted value is determined by the average of all bootstrap samples. This is formalised in equation (15):

$$\hat{f}_{all}(\boldsymbol{X}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(\boldsymbol{X}) \tag{15}$$

Secondly, only a restricted number of variables m is considered at each split. Although m remains the same per split, the variables that are used change. A pre-set number of m variables is randomly selected for every split. By doing so, the splitting conditions differ per tree, which reduces correlation in the distinct trees. The number of variables randomly

selected at each split is a tuneable hyperparameter. There is some disagreement in the existing literature on whether the value of this hyperparameter influences the results of the Random Forest. Breiman (2001) states that although the number of variables divided by three is a good estimate for regression problems, the number of variables divided by two and the number of variables multiplied by two should both be considered. At the same time, Liaw and Wiener (2002) empirically show that the number of variables considered per split does not affect the results of a Random Forest.

The algorithm uses multiple decision trees for its prediction. The number of trees that are used is a hyperparameter. When using too few trees, variance in the data set is likely to be unexplained. On the other hand, adding too many trees leads to high computational costs. Due to the trade-off between prediction accuracy and computational costs, tests on 29 distinct data sets show that the optimal number of trees in a Random Forest is between 64 and 128 (Oshiro, Perez, & Baranauskas, 2012). To be on the safe side, this study uses 128 trees in a Random Forest. Since the Random Forest algorithm is an ensemble method, adding more trees does not increase the risk of overfitting the model (Breiman, 2001).

4.1.3. LASSO-regularised Linear Regression

The main idea of a LASSO-regularised Linear Regression is that a penalty term is applied to the coefficients of an Ordinary Least Squares (OLS) Regression. This penalty term shrinks the coefficients towards zero and, by doing so, prevents overfitting. Adding many variables to a model results in a highly complex model, which is likely to result in excellent results on the training data. However, as it is too specific for the training data, it performs poorly on unseen test data. Therefore, shrinking the coefficients towards zero and reducing the complexity of the model decreases the risk of overfitting.

The size of the penalty term is determined by the penalty parameter λ. This is a tuneable hyperparameter under the constraint that it is positive. The penalty term equals λ times the sum of all coefficients. This means that no penalty term exists if $\lambda = 0$ and that the model is simply an OLS Regression. When λ increases, the size of the model's coefficients starts to decrease. All coefficients are equal to zero if $\lambda \to \infty$. The resulting loss function of a LASSO-regularised Linear Regression is mathematically depicted in equation (16):

$$L(b_1, \dots, b_m) = \sum_{i=1}^{N} (y_i - \sum_{j=1}^{m} x_{ij} b_j)^2 + \lambda \sum_{j=1}^{m} |b_j| \quad (16)$$

In the equation, b_j represents an unknown regression weight for variable $j = 1, \dots, m$. Also, x_{ij} is an element of the $n \times m$ predictor variable matrix $\boldsymbol{X}$ and y_i is the value of the dependent variable for $i = 1, \dots, n$.

4.2. Dependent and independent variables

The dependent variable is the same for all models, namely the overall performance of a player in year *t*. The independent variables, however, differ per model. Although the same predictors are used for the models, their lag is different. All models use *age*, *height*, *weight*, *preferred position*, *attacking* and *defensive work rate*, and 33 football-specific skills as independent variables. The models that predict performance one year later use the values of these predictors in the year *t – 1*. Although using the same predictors, the three-year predictive models take the values of the year *t – 3*.

4.3. Data transformations

Due to the lag in the independent variables and the way the models are constructed, some data transformations are required. Firstly, the data transformations that are necessary as the model uses data from either one or three years prior will be discussed. As a result of the lag between the dependent and independent variables, several observations were lost. To maintain as much data as possible, two separate data sets are constructed: one data set for the models that predict one year in the future, and one for the models that predict three years in the future. Regarding the former data set, observations of a player that was not in *FIFA* one year earlier are removed. This was necessary to ensure that all independent variables were available for all observations. Afterwards, observations with missing values are removed. These transformations resulted in a data set of 98,240 observations, featuring 25,588 players. This data set is further referred to as D_1.

Almost the same holds for the data set that is required for the model that predicts three years in the future. In this case, however, observations of a player that was not in *FIFA* one year or three years prior are removed. A player is kept in the data set in the unlikely scenario when he is in *FIFA* one and three years prior, but is excluded from the game two years prior. Observations of a player with lacking data for the year before are removed to be able to

compare the predictive performance of models that predict over a different period, while using the same data. After removing observations with missing values, a data set of 52,181 observations containing 15,385 players is left. This data set is a subset of D_1, and is referred to as D_3.

Secondly, three data transformations were required to create the desired models. The first transformation is that both data sets were randomly split into training, validation, and test sets. This means that it is possible for observations about one player to be scattered over the three data sets. The training set consists of 70% of the data, while both others contain 15% of the data. The second transformation has to be carried out due to ANNs not being able to handle categorical variables. Therefore, all categorical variables have been one-hot encoded before an ANN is constructed. This means that for every category of a categorical variable, a new variable is created. The variable takes the value one if an observation belongs to that category, and the value zero if it does not. One category of every categorical variable is removed to avoid multicollinearity when constructing the ANNs. The standard categorical variables have been used to construct the Random Forests and Linear Regressions. Lastly, it has to be ensured that all independent variables are on the same scale. Otherwise, the models will prioritise variables with high values over variables with smaller values. Therefore, all continuous football-related independent variables are divided by 100. Since these variables were already on a scale from zero to 100, dividing by 100 brings them on a scale from zero to one. This eases the interpretation of the results compared to when these variables would have been normalized. Additionally, other continuous independent variables are normalised to ensure that they are on a scale from zero to one as well. As these variables are on a different range than the football-related variables, simply dividing by 100 would yield erroneous results.

4.4. Hyperparameter optimisation

As mentioned before, the performance of three separate types of machine learning methods are evaluated in this study: an ANN, Random Forest, and LASSO-regularised Linear Regression. All three methods required some hyperparameters to be determined. The number of neurons in a hidden layer, dropout rate, number of epochs, and batch size have to be set before the learning process of an ANN can start. Also, it is tested whether or not the number of variables used at each split affects the results of a Random Forest, due to the

disagreement in the existing literature on this subject. Regarding the LASSO-regularised Linear Regression, the penalty parameter λ has to be determined.

4.4.1. Grid Search method

A straightforward approach into choosing the optimal hyperparameters for a specific data set is the Grid Search method. This means that a broad range is set for every hyperparameter and that all possible combinations of hyperparameters within their pre-set range are considered. This method is used to determine the optimal hyperparameters of the Random Forests and the LASSO-regularised Linear Regressions, as the number of possible combinations is relatively small. Regarding the Random Forest, the possible values of the number of variables used at each is split are those proposed by Breiman (2001). Therefore, the number of variables divided by six, the number of variables divided by three, and the number of variables divided by 1.5 are tried. This means that the values thirteen, seven, and twenty-six are considered as the number of variables used per split. The possible values of the penalty parameter λ in the LASSO-regularised Linear Regressions are 10^s, in which s varies between -5 and 5, with incremental steps of 0.2. The accuracy of the model for every value is calculated based on the validation set, measured in root mean squared error (RMSE). The hyperparameter values with the lowest RMSE are used to train the eventual model.

4.4.2. Bayesian Hyperparameter Optimisation

Considering the many hyperparameters of an ANN, applying the Grid Search method to these models would be a highly time-consuming process. Almost 10,000 combinations of hyperparameters would have to be tried to derive the best combination. A method that largely overcomes the time-complexity problem of the Grid Search method is the Random Search method. This method relies on the same principle as the Grid Search method, namely that many combinations within a pre-set range are tested. The difference is that the Random Search method only tries a certain percentage of the combinations, while the Grid Search method tests all combinations. A downside of this approach, however, is that it does not use the knowledge obtained from prior attempts for choosing its next combination of hyperparameters. Therefore, Bayesian Hyperparameter Optimisation (BHO) as proposed by Snoek, Larochelle and Adams (2012) is applied to tune the hyperparameters. The main difference between the Random Search method and BHO is that the latter uses knowledge obtained from prior efforts to choose its next combination of hyperparameters. BHO

continuously calculates the probability of an improvement of the accuracy given a certain set of hyperparameters. This probability is updated after every iteration. In this article, sequential model-based hyperparameter optimisation is used. This means that trials are run one after another while trying better hyperparameters each time by updating the probability model through Bayesian reasoning.

BHO is mainly based on the surrogate function, which can be defined as $P(y|z)$ (Snoek, Larochelle, & Adams, 2012). In this function, y is binary and indicates whether there is an improvement of the accuracy and z is a set of hyperparameters. This function is responsible for calculating the probability of a set of hyperparameters resulting in higher accuracy. This is schematically visualised in Figure 8. In Figure 8a, only two sets of hyperparameters have been evaluated and it is visible that the black line of the surrogate function is quite far away from the true value. However, Figure 8b shows that after eight evaluations, the surrogate function matches the true value closely. This illustrates how BHO can calculate the optimal combination of hyperparameters.

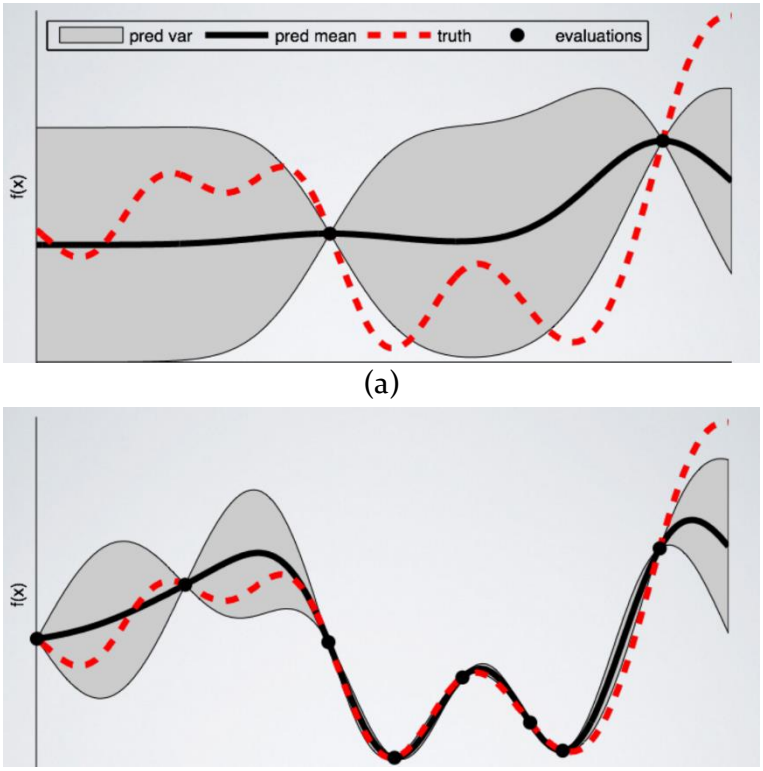


Figure 8: Schematic representation of Bayesian Hyperparameter Optimisation

The surrogate function is calculated through a Gaussian process (Frazier, 2018). The optimisation problem is summarised by an objective function f. The values $z_1, \dots, z_k \in \mathbb{R}$ of f, in which k equals the number of prior tries, are used to identify the optimal set of hyperparameters. This results in a vector of values $[f(z_1), \dots, f(z_k)]$ or shortened as $\boldsymbol{f}(z_{1:k})$. For every z_i, the mean vector μ_i is calculated. Additionally, a covariance matrix $\sum_i$ for each pair of point z_i, z_j is determined through a Gaussian kernel (Rasmussen, 2004). This causes points z_i, z_j that are close in the input space to have a large positive correlation. This represents that these points have more similar value functions than points that are further away from one another in the input space. The resulting distribution of $\boldsymbol{f}(z_{1:k})$ is shown in equation (17):

$$\boldsymbol{f}(z_{1:k}) \sim Normal(\mu(z_{1:k}), \sum(z_i, z)) \tag{17}$$

All values of $\boldsymbol{f}(z_{1:k})$ are known and the aim is to calculate $f(z)$ for a new point z. To do so, it is assumed that $k = n + 1$ and $z_k = z$, indicating that (17) is still valid for $[\boldsymbol{f}(z_{1:k}), f(z)]$. Then, Bayes' rule can be used to determine the conditional distribution of $f(z)$, as visualised in equation (18):

$$\begin{gathered} f(z)|\boldsymbol{f}(z_{1:n}) \sim Normal(\mu_n(z), \sigma_n^2(z)) \\ \mu_n(z) = \sum(z, z_{1:n})\sum(z_{1:n}, z_{1:n})^{-1}\, \boldsymbol{f}(z_{1:n}) - \mu(z_{1:n}) + \mu(z) \\ \sigma_n^2 = \sum(z, z) - \sum(z, z_{1:n})\sum(z_{1:n}, z_{1:n})^{-1}\sum(z_{1:n}, z) \end{gathered} \tag{18}$$

Afterwards, the selection function determines which set of hyperparameters will be tried next. The criteria for this is expected improvement, as proposed by Močkus (1975) and later modified by Jones, Schonlau and Welch (1998). This means that the set of hyperparameters that is expected to result in the highest improvement is chosen for the next try. The selection function is visualised in equation (19):

$$EI_n(z) = [\Delta_n(z)]^+ + \sigma_n(z)\varphi\left(\frac{\Delta_n(z)}{\sigma_n(z)}\right) - |\Delta_n(z)|\varphi\left(\frac{\Delta_n(z)}{\sigma_n(z)}\right) \tag{19}$$

In this equation, $\Delta_n(z) = \mu_n(z) - f_n^*$, which equals the difference between the accuracy of the proposed set of parameters z and the previous best set. This equation shows that a trade-off is present in the algorithm between high improvement in accuracy, depicted by $\Delta_n(z)$, and high uncertainty, depicted by $\sigma_n(z)$.

For this study, 30 different combinations of hyperparameters are tried to determine the optimal set. To maximise accuracy while avoiding large computational costs, the number of

hidden layers is held constant at three. Furthermore, the ReLU activation function is always applied. The best learning rate per model is determined by the Adam optimiser. The chosen ranges of the remaining hyperparameters of the ANNs are depicted in Table 1. The algorithm stops evaluating a combination when the value of the loss function has not improved for 50 epochs, even when it has not completed its pre-set number of epochs. The epoch from which no further improvement was detected, is considered the optimal number of epochs. The maximum batch size is equal to the total number of observations in the training set.

Table 1: Ranges of the ANN hyperparameters

Hyperparameter	Possible values
Number of hidden layers	3
Activation function	ReLU
Number of neurons	$i \in \mathbb{Z}: i \in [3,200]$
Dropout rate	$i \in \mathbb{R}: i \in [0.00001, 0.5]$
Number of epochs	$i \in \mathbb{Z}: i \in [50, 8{,}000]$
Batch size	$i \in \mathbb{Z}: i \in [10{,}000, \max(batch\ size)]$

4.5. Model performance evaluation

In the end, it has to be determined which model predicts the future performance of football players best. Firstly, it has been reviewed whether the expectation that the long-term model is less accurate than the short-term can be confirmed. This is determined by comparing the prediction accuracy of the short-term and long-term models for the three types of models. To avoid any bias due to inequalities between the two data sets, the long-term model is compared to a short-term model that is constructed with the same data. Therefore, one-year models are constructed based on both data set D_1 and D_3. The model based on data set D_1 is considered the predictive model, whereas the model based on data set D_3 is considered the comparative model. The prediction accuracy is measured in RMSE on the test set. When the long-term models have a higher RMSE than the short-term models, it is confirmed that the long-term models are less accurate. Additionally, it has to be determined which of the three types of methods predicts future performance most accurately. This is done by comparing the prediction accuracy of the three types of methods, both for the short-term and the long-term. The model with the lowest RMSE is considered to predict the future performance of football players best. The model with the best prediction performance per prediction period is used for interpretation purposes.

4.6. Model interpretation

A disadvantage of black-box methods, of which ANN and Random Forest are examples, is that the interpretation of the models is less straightforward than it is for less complicated models. Due to the sheer size of black-box methods, it is hard to fully understand what happens between the input and output of the model. The two prediction periods are interpreted separately. Only the best performing model per prediction period is interpreted. Two distinct methods are used to interpret the black-box methods in this research. The test set is used to compile the interpretation of the models.

Firstly, Partial Dependence Plots (PDPs), as suggested by Greenwell, Boehmke and McCarthy (2018), are used to measure the effect of one or more independent variables on the dependent variable. To construct these PDPs, the original training data is copied and the values of variable x_1 are replaced with the constant x_{1i}, in which $x_{11}, x_{12}, \dots, x_{1k}$ are unique predictor values. Then, the average of the predicted values $\overline{\boldsymbol{f}_1}(x_{1i})$ is calculated. The PDP for variable x_1 is constructed by plotting $\{x_{1i}, \overline{\boldsymbol{f}_1}(x_{1i})\}$ for $i = 1,2, \dots, k$. If a second variable is added to the PDP, not only the values of x_1 but also those of x_2 are replaced. The average of $\overline{\boldsymbol{f}_{1,2}}(x_{1i}, x_{2i})$ is calculated for every combination of x_{1i} and x_{2i}. This is visualised in a plot using a colour scale representing $\overline{\boldsymbol{f}_{1,2}}(x_{1i}, x_{2i})$, and with x_{1i} on the horizontal axis and x_{2i} on the vertical axis. Similarly, more variables can be added to the PDP, although visualising the results can become troublesome due to the requirement of more than two dimensions in a visualisation.

Secondly, a Variable Importance Plot (VIP) containing Permutation Feature Importance Scores (PFISs), as suggested by Breiman (2001), is constructed to determine which variables contribute most to the prediction. It is recognised that the future performance of goalkeepers and non-goalkeepers is likely to be dependent on different variables. Future performance of goalkeepers is expected to be determined by goalkeeping-related skills, while the opposite is true for non-goalkeepers. Therefore, separate VIPs are created for goalkeeping and non-goalkeeping skills. Only goalkeepers in the test set are taken into account when constructing the goalkeeping VIPs, while only non-goalkeepers are used for the non-goalkeeping VIPs. To calculate the PFISs, a baseline of the model accuracy s on the original data set measured in RMSE is taken. Then, the values of variable x_1 are randomly shuffled, creating data set D_{x_1}. The model accuracy s_{x_1} is calculated again, but now by using

data set D_{x_1}. The difference between s and s_{x_1} is the PFIS i_{x_1} of variable x_1. This is repeated for all variables.

5. Results

This section is divided into three subsections. Firstly, the optimal hyperparameters for the models will be presented. Then, it will be determined which models yield the best results per prediction period. Lastly, the best performing model per prediction period will be interpreted.

5.1. Hyperparameter optimisation

Firstly, the hyperparameters of the models are determined. The hyperparameters of the ANNs are determined through Bayesian Hyperparameter Optimisation. The Grid Search method is applied to optimise the hyperparameters of the Random Forests and LASSO-regularised Linear Regressions. The optimal hyperparameters are depicted in Table 2. These values are used to construct the distinct models.

Table 2: Optimal hyperparameter values

	One-year predictive model	Three-year predictive model	One-year comparative model
Data set	D_1	D_3	D_3
Training observations	68,768	36,526	36,526
Validation observations	14,736	7,828	7,828
Artificial Neural Network			
Number of neurons	174	40	127
Dropout rate	0.00001	0.2512	0.00001
Number of epochs	1,139	76	1,991
Batch size	19,749	6,428	25,034
Random Forest			
Number of variables per split	7	13	7
LASSO-regularised Linear Regression			
Penalty parameter	$10^{-2.6}$	$10^{-2.8}$	$10^{-2.8}$

5.2. Model performance evaluation

As the models can now be constructed, it is determined how the predictive performance of the ANNs compares to the predictive performance of the Random Forests and the LASSO-regularised Linear Regressions. In addition to this, the accuracy of the different prediction

periods is compared. The RMSE of a model is used as comparison measure. Whereas the models are trained on the training and validation set, the accuracy of the model is calculated based on the test set. Table 3 shows the predictive performance and computational cost per model. The computational costs are based on the usage of a 2.40 GHz 2397 MHz dual-core CPU. Figure 9 shows a scatterplot of the fit of the one-year predictive ANN. Scatterplots showing the fit of the other one-year predictive models are depicted in Appendix C.

Table 3: Comparison of the different models

	One-year predictive model		Three-year predictive model		One-year comparative model	
	RMSE	Training time	RMSE	Training time	RMSE	Training time
Artificial Neural Network	2.742	47m 39s	5.985	1m 59s	2.418	37m 44s
Random Forest	2.811	105m 53s	3.441	22m 23s	2.445	13m 15s
LASSO-regularised Linear Regression	3.806	< 1s	4.459	< 1s	3.548	< 1s

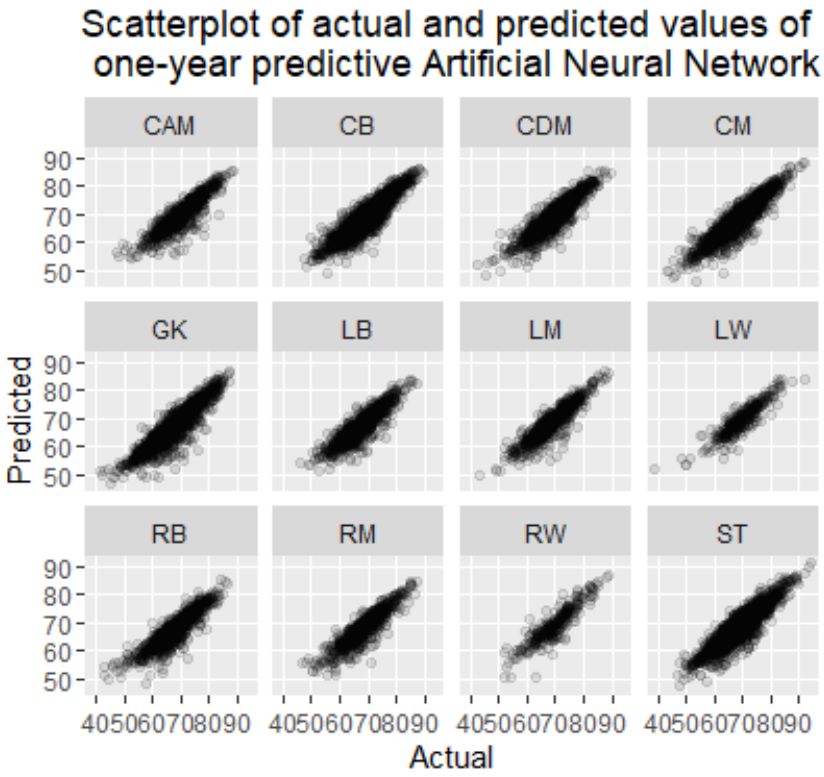


Figure 9: Scatterplot of actual and predicted values per position of the one-year predictive Artificial Neural Network

Based on Table 3, Figure 9, and Figure 16 and Figure 17 in Appendix C, it can be stated that the ANN performs best when predicting performance in one year. It must be noted, however, that the difference in prediction accuracy between the ANN and the Random

Forest is small for the one-year predictive model. Despite their similarities in prediction accuracy, the Random Forest takes more than double the training time to achieve this. It is visible in Figure 17 in Appendix C that the LASSO-regularised Linear Regression predicts relatively similar to the other two models regarding moderate players. When it comes to the most extreme values, however, it is visible that the other two models achieve a higher accuracy. The LASSO-regularised Linear Regression appears to overestimate the future performance of low-skilled players and to underestimate the future level of high-skilled players. As only the best performing model will be interpreted, the ANN one-year predictive model will be chosen for this purpose.

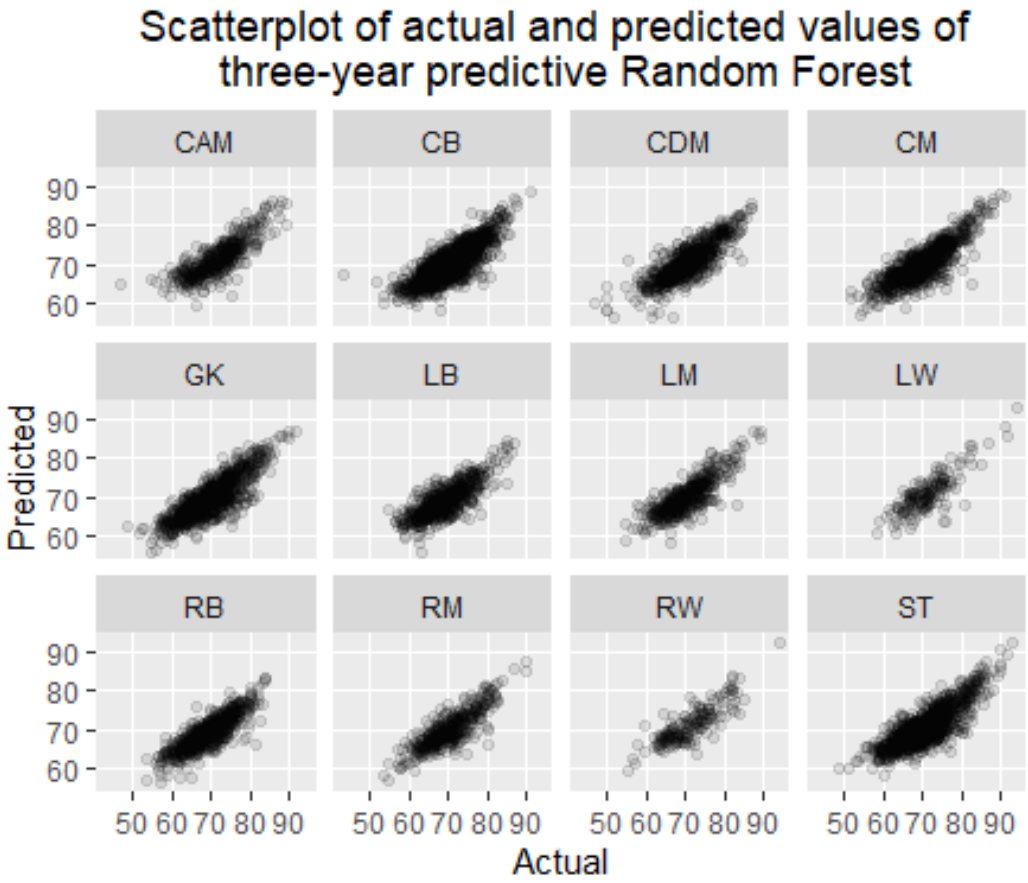


Figure 10: Scatterplot of actual and predicted values per position of the three-year predictive Random Forest

Figure 10 shows a scatterplot of the predicted and actual values of the three-year Random Forest. Based on Table 3, Figure 10, and Figure 18 and Figure 19 in Appendix C, it can be stated that the Random Forest predicts future performance best of the three-year predictive model. In this case, it seems that the ANN is simply too inaccurate for all values. The longer training time of the Random Forest is justified by its better performance. The LASSO-

regularised Linear Regression appears to suffer from the same problem as for the one-year predictive model: it predicts similar to the Random Forest for moderate values, but the Random Forest predicts better for extreme values. However, it is notable that the LASSO-regularised Linear Regression performs better than the ANN, especially when considering the difference in training time. In the end, the Random Forest is used for interpretation.

Additionally, a comparison between the prediction periods is made. To do so, the performance of the three-year predictive model and the one-year comparative model is compared. This is done to ensure that the same data set, namely data set D_3, is used for both models in the comparison. Table 3 shows that the RMSE of the one-year comparative model is substantially lower than the RMSE of the three-year predictive model for all three methods. Furthermore, an important notion is that the RMSE of the one-year comparative models is also lower than the RMSE of the one-year predictive models. This seems counterintuitive: the one-year predictive model is trained on more observations than the one-year comparative model and is, therefore, expected to predict more accurately. A possible explanation for this lies in the difficulty of the data. Data set D_1, which is used for the one-year predictive model, contains all players that were in *FIFA* for two consecutive years. Players were required to be present in the game for four consecutive years in order to be included in data set D_3, which is used for the one-year comparative model. Figure 11 shows a scatterplot of the actual and predicted values of the one-year comparative ANN. The actual and predicted values of the one-year predictive model are depicted in Figure 9. It is visible that data set D_3 has fewer players with an overall score below 55 than data set D_1. There are only 28 players with an overall score below 55 in data set D_3, whereas there are 330 players scoring lower than 55 in data set D_1. It is likely that players present in data set D_1 but not in data set D_3 have a relatively low overall score. Otherwise, they would have continued their professional football career and would be present in *FIFA* for a longer period. Both models overestimate the future performance of these low-skilled players. However, as the one-year predictive model has to predict more low-skilled players, it has a more profound influence on its RMSE than the three-year predictive model. This potentially explains why the one-year comparative model achieved a higher predictive performance using less data.

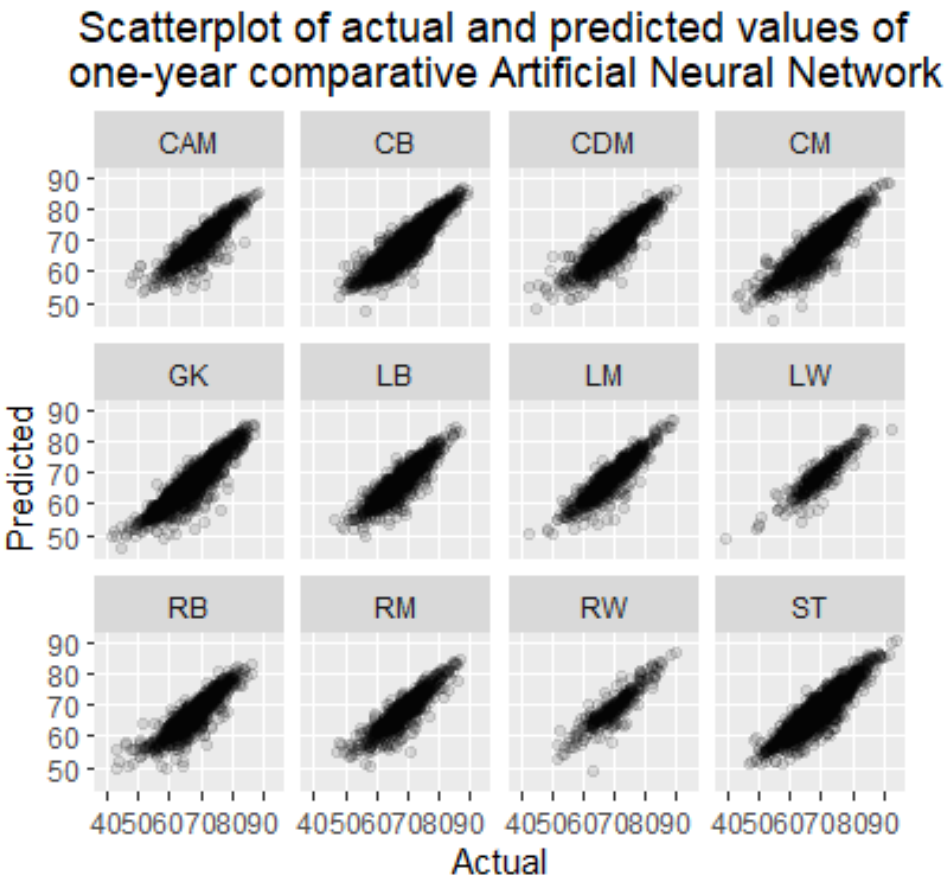


Figure 11: Scatterplot of actual and predicted values of the one-year comparative Artificial Neural Network

5.3. Model interpretation

As the performance of the models has now been evaluated, the model with the best performance per prediction period will be interpreted. Therefore, the ANN is interpreted for the prediction period of one year, whereas the Random Forest is interpreted for the prediction period of three years.

5.3.1. One-year prediction

PDPs are constructed for all variables in the one-year predictive ANN. All PDPs are depicted in Appendix D. It is visible that *short passing*, *dribbling*, *ball control*, *reactions*, *standing tackle* and *sliding tackle* are all strongly positively related to overall performance in the upcoming year. This indicates that players who perform well regarding these skills are likely to achieve higher levels of overall performance in the next year than players who perform worse on these skills. All other football-related variables are either positively related or unrelated to future performance. The only negative relationship observed is between *age* and future performance. This means that a player is likely to obtain a lower overall level in

the upcoming year than a younger player. According to the PDP of *age*, the age-related decay of a player starts from the beginning. *Weight*, *height*, and *preferred foot* appear to be unrelated to future performance. Lastly, there is some variety visible regarding the average predicted performance depending on the *player position*. Goalkeepers are expected to perform best on an average level, whereas centre backs perform worst compared to players on other positions. This difference does not result in a disparity of the prediction performance between player positions. Table 4 shows that the RMSE per player position is approximately equal.

Table 4: Predictive performance of one-year predictive Artificial Neural Network per player position

Position		RMSE	Observations
Centre attacking midfielder	CAM	2.957	864
Centre back	CB	2.539	2,592
Centre defensive midfielder	CDM	2.775	1,286
Centre midfielder	CM	2.783	1,628
Goalkeeper	GK	2.886	1,715
Left back	LB	2.675	1,110
Left midfielder	LM	2.641	787
Left winger	LW	2.797	323
Right back	RB	2.755	1,165
Right midfielder	RM	2.787	783
Right winger	RW	2.865	296
Striker	ST	2.740	2,187

The previously discussed PDPs only represent the main effect of a variable on future performance. Possible interaction effects, however, have not been studied yet. In addition to the main effect of football skills on future performance, it is also investigated how the relationship between these two differs per *player position* and *age*. Firstly, PDPs are created for all variables in the one-year predictive model, which visualise the interaction effect of player position. These PDPs are shown in Appendix E. Mixed results are found from the PDPs, as shown in Figure 12. It must be noted that the predicted performance levels are also influenced by the main effect of player position and the specific skill. The PDPs of some variables show that the relationship between future performance and a football skill differ per player position. On the left side of Figure 12, it is visible that *heading accuracy* has more impact on the future performance of right backs than centre backs, for example. However, most PDPs show that the relationship is fairly similar for the different player positions. This is shown on the right side of Figure 12, where the marginal effect of an increase in *positioning* skills is approximately the same for all positions.

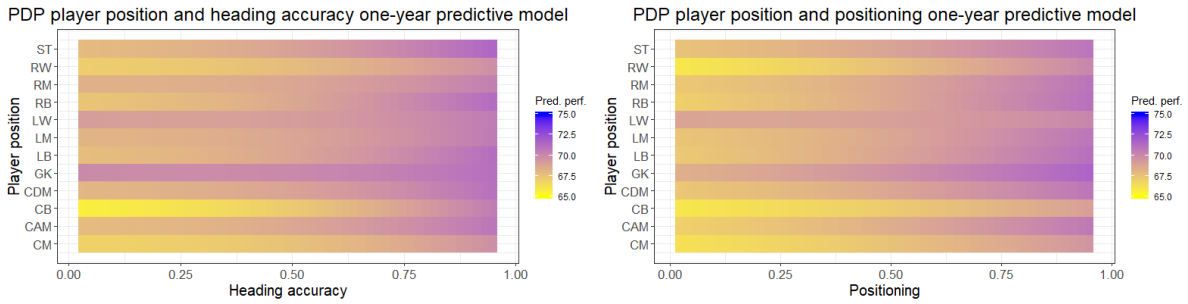


Figure 12: Comparison of Partial Dependence Plots of a variable with a clear interaction effect with player position (left) and without (right) based on the one-year predictive Artificial Neural Network

Secondly, the same has been performed for the interaction effect of *age*. The PDPs are depicted in Appendix F. In general, most PDPs show the effect as depicted in Figure 13. The combination of high age and low performance level of a skill leads to a low predicted performance level, compared to a younger player that performs better in that skill. It is important to note, however, that this is likely to be the result of the main effects of *age* and the particular skill. *Age* is negatively related to future performance on its own, while skill levels are positively related to this factor. As a result, it appears that there is no interaction between age and skill levels when it comes to their relationship with future performance.

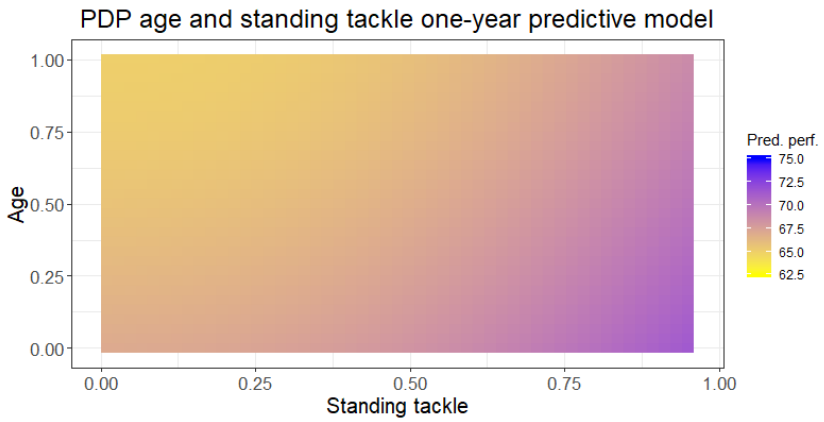


Figure 13: Partial Dependence Plot of standing tackle and age based on the one-year predictive Artificial Neural Network

Besides PDPs, a VIP of the permutation scores is constructed to determine which variables influence the predictions most. Separate VIPs are constructed for goalkeeping and non-goalkeeping skills. Only goalkeepers are used to identify the most important goalkeeping skills, while only non-goalkeepers are used to determine the most important non-goalkeeping skills. The VIP containing the 25 most important non-goalkeeping skills is depicted on the left of Figure 14, whereas the VIP of the goalkeeping skills is shown on the

right. On the left side of Figure 14, it is visible that randomly shuffling the data of both sorts of tackles has the strongest impact on the RMSE of the model. This indicates that these two variables contribute most to the predictions. It is depicted that being a centre back, and *finishing* and *marking* skills are important predictors as well. On the right side of Figure 14, it can be seen that *diving*, *reflexes*, *positioning*, and *handling* are almost equally important for the predictions of goalkeepers. It can be stated that the *kicking* skills of a goalkeeper contribute little to the model.

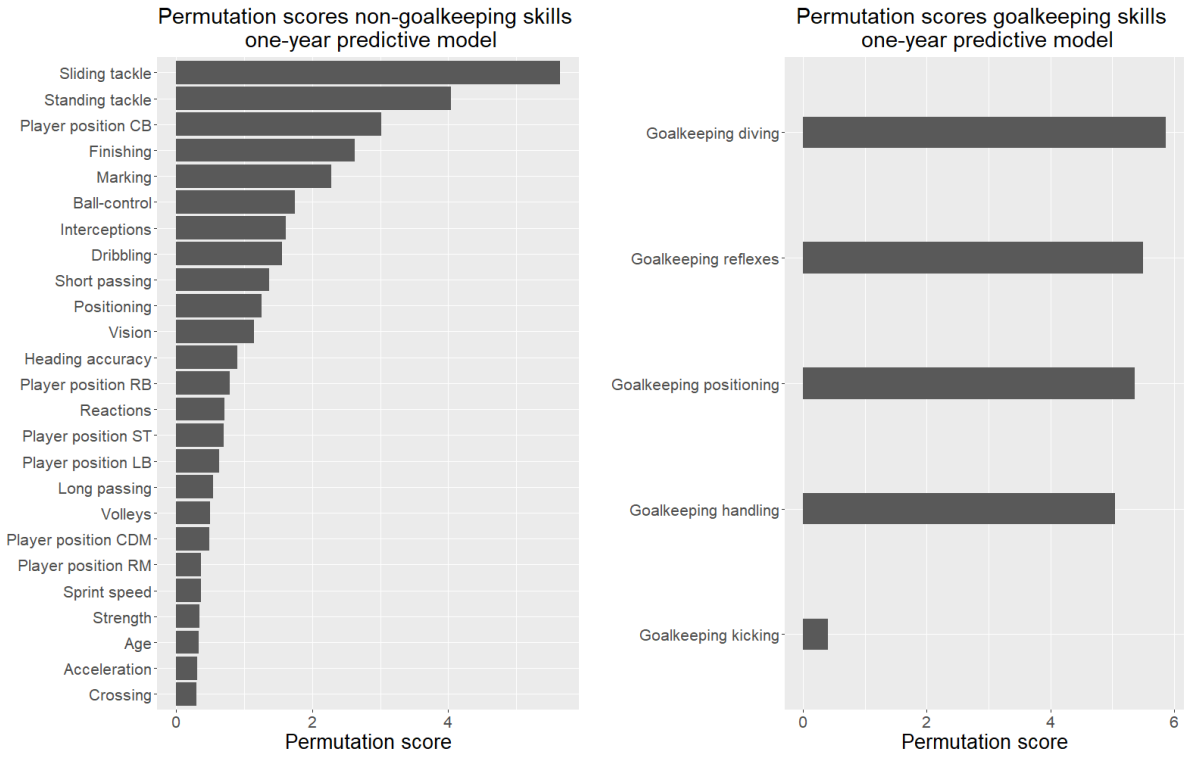


Figure 14: Permutation scores of the 25 most important non-goalkeeping variables (left) and the goalkeeping variables (right) of the one-year predictive Artificial Neural Network

5.3.2. Three-year prediction

As the Random Forest performs best on a three-year prediction period, this model is used for interpretation purposes. All steps taken to interpret the one-year predictive ANN are repeated for the three-year predictive Random Forest. The PDPs visualising the main effect of a variable on future performance are shown in Appendix G. Again, it is visible that all football skills are positively related or unrelated to future performance. The variables with the strongest relationship in the three-year predictive Random Forest are partly similar to those in the one-year predictive ANN. The PDPs show that *ball control*, *reactions*, *vision*, *standing tackle* and *sliding tackle* possess the strongest relationship with future

performance. *Passing* and *dribbling*, which were both found to be strongly related to future performance in the one-year predictive ANN, also show strong positive relationships with future performance in the three-year predictive Random Forest. A clear negative relationship is present between a player's *age* and his future performance. The PDP is gradually sloping down from a normalised age of 0.25, which equals an actual age of approximately 23. *Height*, *weight*, and *preferred foot* appear to be unrelated to future performance. As for the one-year predictive ANN, some variety in predicted performance is visible per *player position*. However, this time goalkeepers and centre backs are expected to perform best. This is in contrast with the one-year predictive ANN, which showed that centre backs are expected to perform worst. Nevertheless, no major differences in prediction accuracy per position are found, as shown in Table 5.

Table 5: Predictive performance of three-year predictive Random Forest per player position

Position		RMSE	Observations
Centre attacking midfielder	CAM	3.607	422
Centre back	CB	3.364	1,402
Centre defensive midfielder	CDM	3.460	695
Centre midfielder	CM	3.528	821
Goalkeeper	GK	3.594	893
Left back	LB	3.434	611
Left midfielder	LM	3.316	439
Left winger	LW	3.296	161
Right back	RB	3.095	632
Right midfielder	RM	3.273	392
Right winger	RW	3.536	175
Striker	ST	3.564	1,185

As for the one-year predictive ANN, PDPs visualising the interaction effects of *player position* and *age* are created. These PDPs are shown in Appendix H and I, respectively. The PDPs visualising the interaction effect of *player position* all look highly similar. The relationship between football skills and future performance is the same for all player positions. As a result, it can be stated that there is no interaction effect of *player position* when it comes to predicting performance in three years. More variety is visible in the PDPs depicting the interaction effect of *age*. However, the extreme values shown in the PDPs can be explained by a combination of two separate main effects, as is the case for the one-year predictive ANN. Therefore, there is also no interaction effect of *age* on the relationship between football skills and performance in three years.

Again, two separate VIPs are created for goalkeeping and non-goalkeeping skills. The VIP of the non-goalkeeping skills is shown on the left side of Figure 15, while the VIP of the goalkeeping skills is depicted on the right side of that figure. It is visible that, as in the one-year predictive ANN, *sliding tackle* is the most important predictor in the model. For the three-year predictive Random Forest, however, *standing tackle* is less important than for the one-year predictive ANN. When considering the three-year predictive Random Forest, *ball control*, *reactions*, and *vision* skills appear to be more important. On the right side of Figure 15, it can be seen that *reflexes* contribute most to the prediction of performance in three years. The difference in permutation scores between the other skills is small.

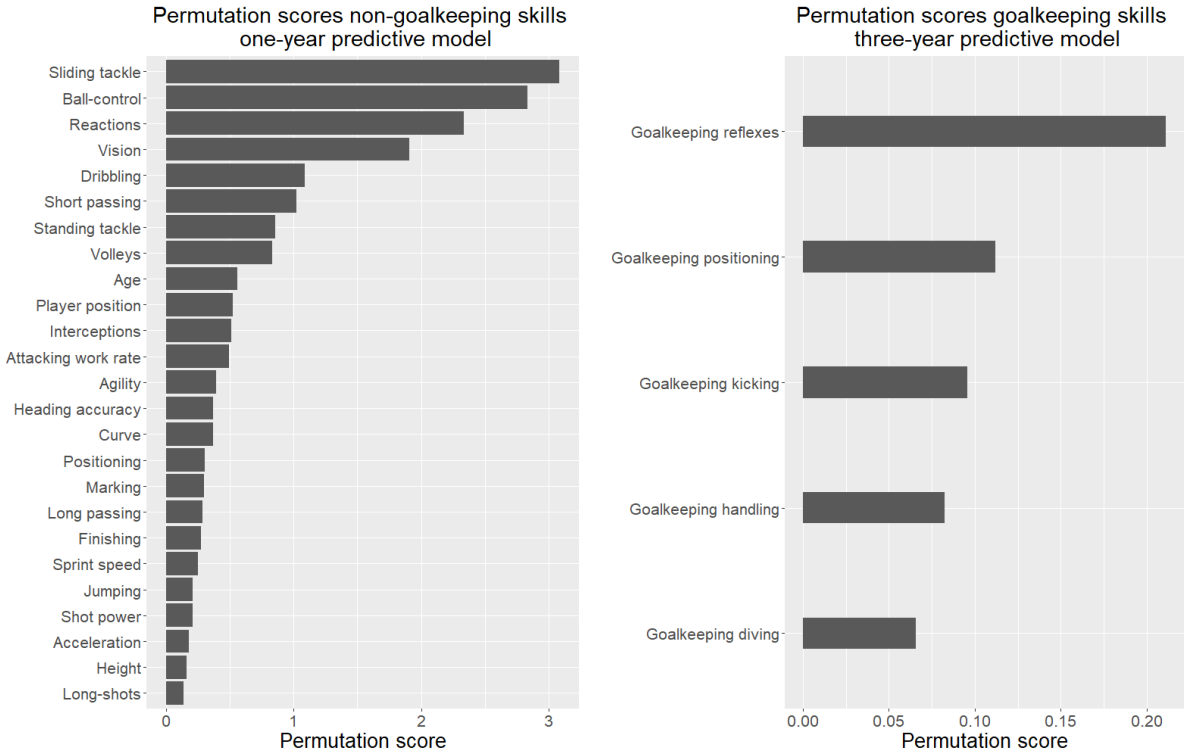


Figure 15: Permutation scores of the 25 most important non-goalkeeping variables (left) and the goalkeeping variables (right) of the three-year predictive Random Forest

6. Discussion

At the beginning of this study, five subquestions have been presented that build towards answering the main research question. In this section, these subquestions will be answered in order and compared to the existing literature.

The first subquestion asks how player characteristics are related to future performance. It can be stated that all football skills are either positively related or unrelated to future performance. No football skills were found to be negatively related to future performance. It was recognised, however, that age is negatively related to future performance. This negative relationship is present for all ages in the one-year predictive ANN. Players were found to lose approximately one point every seven years, under the assumption that all other variables are held constant. This seemingly contradicts the research by Schulz and Curnow (1988), which states that the peak performance age of football players is 25. At the same time, however, it appears to confirm the statement of Schroeder and Salthouse (2004) that mental decline starts in a person's early 20's. Despite this, it was expected that the positive effect of more physical strength would be stronger than the negative effect of mental decline. Nevertheless, this is not confirmed by the results. The three-year predictive Random Forest, on the other hand, shows that the decline in performance only starts at an age of 25. From this point, players lose approximately half a point every seven years. This results appears to confirm both the peak performance age as stated by Schulz and Curnow (1988) and the starting age of mental decline as stated by Schroeder and Salthouse (2004).

The second subquestion examines how the importance of football skills differs when the prediction period changes. In this study, two prediction periods were taken into consideration: one year and three years. As goalkeeping and non-goalkeeping skills are deemed too different to be compared, the interpretation of these two types of skills was split. Considering the one-year prediction period, it was found that standing tackle, sliding tackle, finishing, and marking are the non-goalkeeping football skills that influence the predictions most. Regarding the goalkeeping skills, it can be stated that kicking has a marginal effect on the predictions. The other goalkeeping skills (i.e., diving, positioning, reflexes, and handling) were almost equally important for the predictions. When looking at the three-year prediction period, it can be stated that sliding tackle, ball control, reactions, and vision are the most important predictors of future performance. Reflexes were the most influential goalkeeping skills. Most of these skills (e.g., tackling, finishing, reflexes, and ball control) can be considered motor skills, whereas marking and positioning are more related to perceptual skills, for example. As a result of this, it is hard to connect these findings to existing literature. Literature stated that motor skills are the most important predictors of future performance. In that case, it would be expected that all motor skills are important

determinants of future performance. Nevertheless, it is found that only some motor and perceptual skills are important when predicting future performance. It is unclear why this division in importance exists. It appears, however, that future performance is not necessarily dependent on motor, perceptual, and cognitive skills, but on distinct football skills.

The third and fourth subquestions are about the influence of age and player position on the relationship between player characteristics and future performance. The results showed no interaction effect of age. This means that the effect of having a certain level of skills is approximately the same for every age. The results considering the interaction effect of player position were slightly mixed. It was found that the relationships between most football skills and future performance are unaffected by player position. Only a highly limited number of football skills, such as heading accuracy, were influenced by player position. For example, heading accuracy turned out to have a stronger influence on the future performance of centre backs than the future performance of players in other positions. This partly contradicts research by Di Salvo et al. (2007) stating that a different set of skills per position determines a player's performance.

The fifth subquestion asks which machine learning model yields the highest predictive power. When considering the one-year prediction period, the ANN was found to yield the lowest RMSE. It must be noted, however, that the difference with the Random Forest was relatively small. The results showed that the Random Forest yields the lowest RMSE when predicting over a three-year period. This is largely in accordance with the existing literature. Literature about this subject stated that ANNs typically result in the highest predictive power, but that better performance of other methods is possible. This was confirmed by the results of this research.

7. Conclusion

The aim of this study was to answer the research question on what sport-specific player characteristics influence the future performance level of male professional football players. By doing so, the future performance of professional soccer players can be predicted as accurately as possible, and the determinants of the predictions can be identified. For this

purpose, data from the soccer video game *FIFA* was used and multiple predictions methods were compared.

It was found that an ANN yields the best results when predicting on the short-term (i.e., one year), whereas a Random Forest yields the best results on the long-term (i.e., three years). Based on scatterplots depicting the actual and predicted levels of performance, it can be concluded that the predictions are stable and accurate. Therefore, the models are deemed suitable to be used on real player data. As expected, the prediction accuracy decreases as the prediction period increases. Although existing literature suggested that future performance of football players is mainly dependent on motor skills, this was not confirmed by the results of this study. It was found that a small number of individuals football skills are strongly related to future performance. These skills included cognitive and perceptual skills as well. Which skills influence future performance most differed per prediction period.

This study faces some limitations that are mainly related to the data that was used. Firstly, it should be noted that the data was somewhat subjective. Although the data can be deemed reliable due to 6,000 reviewers determining player scores, it is unclear what the actual difference is between players scoring 80 and 90 on sprint speed. Ideally, all players would be tested objectively on all skills listed in the game with a certain performance resulting in a corresponding score. This would be a highly time-consuming process and almost unfeasible in a short time span. However, it can be investigated how accurately video game data matches real player data on a smaller scale. Many sports video games use highly detailed player data. If it turns out that this data represents the real situation fairly accurately, this data can be used for other sport-related research. This is something that has not been investigated yet. Secondly, the data used only contains professional male soccer players. It is unclear if the results are also valid for female or non-professional players. Including these groups would highly increase the study's validity. It is recognised, however, that it is even harder to find reliable data for these types of players. For example, both female and amateur leagues are rarely broadcasted on television, making it difficult to review the quality of the players by a large group of reviewers.

8. References

Aartsen, M. J., Smits, C. H., Van Tilburg, T., Knipscheer, K. C., & Deeg, D. J. (2002). Activity in Older Adults: Cause or Consequence of Cognitive Functioning? A Longitudinal Study on Everyday Activities and Cognitive Performance in Older Adults. *The Journals of Gerontology, 57*(2), 153-162. doi:10.1093/geronb/57.2.P153

Abbott, A., & Collins, D. (2004). Eliminating the dichotomy between theory and practice in talent identification and development: considering the role of psychology. *Journal of Sports Sciences, 22*(5), 395-408. doi:10.1080/02640410410001675324

Ali, A. (2011). Measuring soccer skill performance: a review. *Scandinavian Journal of Medicine & Science in Sports, 21*(2), 170-183. doi:10.1111/j.1600-0838.2010.01256.x

Allen, J. S., Bruss, J., Brown, C. K., & Damasio, H. (2005). Normal neuroanatomical variation due to age: The major lobes and a parcellation of the temporal region. *Neurobiology of Aging, 26*(9), 1245-1260. doi:10.1016/j.neurobiolaging.2005.05.023

Baltes, P. B., & Baltes, M. M. (1990). *Successful Aging: Perspectives from the Behavioral Sciences.* Cambridge, UK: Press Syndicate of the University of Cambridge.

Barnett, L. M., van Beurden, E., Morgan, P. J., Brooks, L. O., & Beard, J. R. (2008). Does Childhood Motor Skill Proficiency Predict Adolescent Fitness? *Medicine & Science in Sports & Exercise, 40*(12), 2137-2144. doi:10.1249/MSS.0b013e31818160d3

Baum, S. M., Owen, S. V., & Oreck, B. A. (1996). Talent Beyond Words: Identification of Potential Talent in Dance and Music in Elementary Students. *Gifted Child Quarterly, 40*(2), 93-101. doi:10.1177/001698629604000206

Björkman, I., Ehrnrooth, M., Mäkelä, K., Smale, A., & Sumelius, J. (2013). Talent or not? Employee reactions to talent identification. *Human Resource Management, 52*(2), 195-214. doi:10.1002/hrm.21525

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140. doi:10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. Retrieved from https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* Boca Raton, FL: CRC Press.

Brouwers, J., de Bosscher, V., & Sotiriadou, P. (2012). An examination of the importance of performances in youth and junior competition as an indicator of later success in tennis. *Sport Management Review, 15*(4), 461-475. doi:10.1016/j.smr.2012.05.002

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics, 15*(1), 27-33. doi:10.1016/j.aci.2017.09.005

Bürgi, F., Meyer, U., Granacher, U., Schindler, C., Marques-Vidal, P., Kriemler, S., & Puder, J. J. (2011). Relationship of physical activity with motor skills, aerobic fitness and body fat in preschool children: a cross-sectional and longitudinal study. *International Journal of Obesity, 35*(7), 937-944. doi:10.1038/ijo.2011.54

Cao, C. (2012). *Sports Data Mining Technology Used in Basketball Outcome Prediction.* Retrieved from https://arrow.tudublin.ie/scschcomdis/39/

Chow, S. M., Henderson, S. E., & Barnett, A. L. (2001). The Movement Assessment Battery for Children: A Comparison of 4-Year-Old to 6-Year-Old Children From Hong Kong and the United States. *American Journal of Occupational Therapy, 55*(1), 55-61. doi:10.5014/ajot.55.1.55

Cliff, D. P., Okely, A. D., Smith, L. M., & McKeen, K. (2009). Relationships between Fundamental Movement Skills and Objectively Measured Physical Activity in Preschool Children. *Human Kinetics Journals, 21*(4), 436-449. doi:10.1123/pes.21.4.436

Das, S. (2020). *Top 10 Most Popular Sports in The World*. Retrieved from https://sportsshow.net/top-10-most-popular-sports-in-the-world/

Deloitte. (2020). *Annual Review of Football Finance 2020.* Retrieved from https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance.html

Di Cagno, A., Battaglia, C., Fiorilli, G., Piazza, M., Giombini, A., Fagnani, F., . . . Pigozzi, F. (2014). Motor Learning as Young Gymnast's Talent Indicator. *Journal of Sports & Medicine, 13*(4), 767-773. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4234945/

Di Salvo, V., Baron, R., Tschan, H., Montero, C., J., F., Bachl, N., & Pigozzi, F. (2007). Performance Characteristics According to Playing Position in Elite Soccer. *International Journal of Sports Medicine, 28*(3), 222-227. doi:10.1055/s-2006-924294

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science, 2*(2), 1-10. doi:10.1080/17461390200072201

Elferink-Gemser, M. T., Jordet, G., Coelho-E-Silva, M. J., & Visscher, C. (2011). The marvels of elite sports: how to get there? *British Journal of Sports Medicine, 45*, 683-684. doi:10.1136/bjsports-2011-090254

Elferink-Gemser, M. T., Visscher, C., Richart, H., & Lemmink, K. A. (2004). Development of the Tactical Skills Inventory for Sports. *Perceptual and Motor Skills, 99*(3), 883-895. doi:10.2466/pms.99.3.883-895

Faulkner, J. A., Davis, C. S., Mendias, C. L., & Brooks, S. V. (2008). The aging of elite male athletes: age-related changes in performance and skeletal muscle structure and

function. *Clinical Journal of Sport Medicine, 18*(6), 501-507. doi:10.1097/JSM.0b013e3181845f1c

Feldhusen, J. F. (1994). Talent Identification and Development in Education (TIDE). *Gifted Education International, 10*(1), 10-15. doi:10.1177/026142949401000103

Fisher, A., Reilly, J. J., Kelly, L. A., Montgomery, C., Williamson, A., Paton, J. Y., & Grant, S. (2005). Fundamental Movement Skills and Habitual Physical Activity in Young Children. *Medicine & Science in Sports & Exercise, 37*(4), 684-688. doi:10.1249/01.MSS.0000159138.48107.7D

Fotenos, A. F., Snyder, A. Z., Girton, L. E., Morris, J. C., & Buckner, R. L. (2005). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology, 64*(6), 1032-1039. doi:10.1212/01.WNL.0000154530.72969.11

Frazier, P. I. (2018). *A Tutorial on Bayesian Optimization.* Retrieved from https://arxiv.org/pdf/1807.02811.pdf

Gerrard, B. (2007). Is the Moneyball Approach Transferable to Complex Invasion Team Sports? *International Journal of Sport Finance, 2*(4), 214-225. Retrieved from https://search.proquest.com/docview/229399553?pq-origsite=gscholar

Goble, C. (2019). *Champions League Is Way Bigger Than the Super Bowl*. Retrieved from https://www.one37pm.com/strength/sports/champions-league-soccer-super-bowl

Gonaus, C., & Müller, E. (2012). Using physiological data to predict future career progression in 14- to 17-year-old Austrian soccer academy players. *Journal of Sports Sciences, 30*(15), 1673-1682. doi:10.1080/02640414.2012.713980

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, MA: MIT Press.

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). *A Simple and Effective Model-Based Variable Importance Measure.* Retrieved from https://arxiv.org/pdf/1805.04755.pdf

Haga, M. (2009). Physical Fitness in Children With High Motor Competence Is Different From That in Children With Low Motor Competence. *Physical Therapy, 89*(10), 1089-1097. doi:10.2522/ptj.20090052

Hakes, J. K., & Sauer, R. D. (2006). An Economic Evaluation of the Moneyball Hypothesis. *Journal of Economic Perspectives, 20*(3), 173-185. doi:10.1257/jep.20.3.173

Hakes, J. K., & Sauer, R. D. (2007). The Moneyball Anomaly and Payroll Efficiency: A Further Investigation. *International Journal of Sport Finance, 2*(4), 177-189. Retrieved from https://s3.amazonaws.com/academia.edu.documents/49918215/The_Moneyball_Anomaly_and_Payroll_Effici20161027-16275-1q5932h.pdf?response-content-disposition=inline%3B%20filename%3DThe_Moneyball_anomaly_and_payroll_effici.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X

Hoare, D. G., & Warr, C. R. (2000). Talent identification and women's soccer: An Australian experience. *Journal of Sports Sciences, 18*(9), 751-758. doi:10.1080/02640410050120122

Hrysomallis, C. (2011). Balance Ability and Athletic Performance. *Sports Medicine, 41*, 221-232. doi:10.2165/11538560-000000000-00000

Iivonen, S., Sääkslahti, A., & Nissinen, K. (2011). The development of fundamental motor skills of four- to five-year-old preschool children and the effects of a preschool physical education curriculum. *Early Child Development and Care, 181*(3), 335-343. doi:10.1080/03004430903387461

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization, 13*, 455-492. doi:10.1023/A:1008306431147

Kahn, J. (2003). *Neural Network Prediction of NFL Football Games.* World Wide Web Electronic Publication. Retrieved from http://homepages.cae.wisc.edu/~ece539/project/f03/kahn.pdf

Kannekens, R., Elferink-Gemser, M. T., & Visscher, C. (2011). Positioning and deciding: key factors for talent development in soccer. *Scandinavian Journal of Medicine & Science in Sports, 21*(6), 846-852. doi:10.1111/j.1600-0838.2010.01104.x

Karsoliya, S. (2012). Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture. *International Journal of Engineering Trends and Technology, 3*(6), 714-717. Retrieved from http://ijettjournal.org/volume-3/issue-6/IJETT-V3I6P206.pdf

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization.* Retrieved from https://arxiv.org/abs/1412.6980

Lange, D. (2019). *Top-20 European soccer clubs by total revenue 2018/19 season*. Retrieved from https://www.statista.com/statistics/271581/revenue-of-soccer-clubs-worldwide/

Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game.* New York, NY: W. W. Norton & Company.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2/3*, 18-22. Retrieved from https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf

Maier, T., Meister, D., Trösch, S., & Wehrlin, J. P. (2018). Predicting biathlon shooting performance using machine learning. *Journal of Sports Sciences, 36*(20), 2333-2339. doi:10.1080/02640414.2018.1455261

Maszczyk, A., Golaś, A., Pietraszewski, P., Roczniok, R., Zając, A., & Stanula, A. (2014). Application of Neural and Regression Models in Sports Results Predictions.

Procedia - Social and Behavioral Sciences, 117, 482-487. doi:10.1016/j.sbspro.2014.02.249

McCabe, A., & Trevethan, J. (2008). Artificial Intelligence in Sports Prediction. *IEEE*, 1194-1197. doi:10.1109/ITNG.2008.203

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*, 115-133. doi:10.1007/BF02478259

McKenzie, T. L., Sallis, J. F., Broyles, S. L., Zive, M. M., Nader, P. R., Berry, C. C., & Brennan, J. J. (2002). Childhood Movement Skills: Predictors of Physical Activity in Anglo American and Mexican American Adolescents? *Research Quarterly for Exercise and Sport, 73*(3), 238-244. doi:10.1080/02701367.2002.10609017

McMorris, T. M., & Graydon, J. (1996). The Effect of Exercise on the Decision-Making Performance of Experienced and Inexperienced Soccer Players. *Research Quarterly for Exercise and Sport, 67*(1), 109-114. doi:10.1080/02701367.1996.10607933

McMorris, T. M., & Graydon, J. (1997). The effect of exercise on cognitive performance in soccer-specific tests. *Journal of Sports Sciences, 15*(5), 459-468. doi:10.1080/026404197367092

Mills, A., Butt, J., Maynard, I., & Harwood, C. (2014). Toward an Understanding of Optimal Development Environments Within Elite English Soccer Academies. *The Sport Psychologist, 28*, 137-150. doi:10.1123/tsp.2013-0018

Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. B. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems, 21*(7), 551-562. doi:10.1016/j.knosys.2008.03.016

Močkus, J. (1975). On Bayesian Methods for Seeking the Extremum. In G. I. Marchuk, *Optimization Techniques IFIP Technical Conference. Lecture Notes in Computer Science* (pp. 400-404). Berlin: Springer.

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning*, 807-814. Retrieved from https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In P. Perner, *Machine Learning and Data Mining in Pattern Recognition* (pp. 154-168). Berlin: Springer.

Plassman, B. L., Welsh, K. A., Helms, M., Brandt, J., Page, W. F., & Breitner, J. C. (1995). Intelligence and education as predictors of cognitive state in late life: A 50-year follow-up. *Neurology, 45*(8), 1446-1450. doi:10.1212/WNL.45.8.1446

Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials, 15*(3), 9-15. doi:10.1109/45.535226

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In O. Bousquet, U. Von Luxburg, & G. Rätsch, *Advanced Lectures on Machine Learning* (pp. 63-71). New York, NY: Springer.

Reilly, T., Williams, A. M., Nevill, A., & Franks, A. (2000). A multidisciplinary approach to talent identification in soccer. *Journal of Sports Sciences, 18*, 695-702. doi:10.1080/02640410050120078

Rochester, N., Holland, J., Haibt, L., & Duda, W. (1956). Tests on a cell assembly theory of the action of the brain, using a large digital computer. *IRE Transactions on Information Theory, 2*(3), 80-93. doi:10.1109/TIT.1956.1056810

Rönnlund, M., Nyberg, L., Bäckman, L., & Nilsson, L.-G. (2005). Stability, Growth, and Decline in Adult Life Span Development of Declarative Memory: Cross-Sectional and Longitudinal Data From a Population-Based Study. *Psychology and Aging, 20*(1), 3-18. doi:10.1037/0882-7974.20.1.3

Rösch, D., Hodgson, R., Peterson, L., Graf-Baumann, T., Junge, A., Chomiak, J., & Dvorak, J. (2000). Assessment and Evaluation of Football Performance. *The American Journal of Sports Medicine, 28*(5), 29-39. doi:10.1177/28.suppl_5.s-29

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning Internal Representations by Error Propagation.* Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a164453.pdf

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536. doi:10.1038/323533a0

Salat, D. H., Buckner, R. L., Snyder, A. Z., Greve, D. N., Desikan, R. S., Busa, E., . . . Fischl, B. (2004). Thinning of the Cerebral Cortex in Aging. *Cerebral Cortex, 14*(7), 721-730. doi:10.1093/cercor/bhh032

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiol Aging, 30*(4), 507-514. doi:10.1016/j.neurobiolaging.2008.09.023

Sanders-Woudstra, J. A., Verhulst, F. C., & De Witte, H. F. (1993). *Leerboek Kinder- en Jeugdpsychiatrie.* Assen: Uitgeverij Koninklijke Van Gorcum.

Schroeder, D. H., & Salthouse, T. A. (2004). Age-related effects on cognition between 20 and 50 years of age. *Personality and Individual Differences, 36*(2), 393-404. doi:10.1016/S0191-8869(03)00104-1

Schulz, R., & Curnow, C. (1988). Peak Performance and Age Among Superathletes: Track and Field, Swimming, Baseball, Tennis, and Golf. *Journal of Gerontology, 43*(5), 113-120. doi:10.1093/geronj/43.5.P113

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in neural information processing systems*, pp. 2951-2959. Retrieved from https://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf

Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L., & Toga, A. W. (2003). Mapping cortical change across the human life span. *Nature Neuroscience, 6*(3), 309-315. doi:10.1038/nn1008

Stodden, D., Langendorfer, S., & Roberton, M. A. (2009). The Association Between Motor Skill Competence and Physical Fitness in Young Adults. *Research Quarterly for Exercise and Sport, 80*(2), 223-229. doi:10.1080/02701367.2009.10599556

Stratton, G., Reilly, T., Williams, A. M., & Richardson, D. (2004). *Youth Soccer: From Science to Performance.* Abingdon: Routledge.

Sullivan, E. V., & Pfefferbaum, A. (2006). Diffusion tensor imaging and aging. *Neuroscience and Biobehavior Reviews, 30*(6), 749-761. doi:10.1016/j.neubiorev.2006.06.002

Unnithan, V., White, J., Georgiou, A., Iga, J., & Drust, B. (2012). Talent identification in youth soccer. *Journal of Sports Sciences, 30*(15), 1719-1726. doi:10.1080/02640414.2012.731515

Williams, A. M., & Davids, K. (1998). Visual Search Strategy, Selective Attention, and Expertise in Soccer. *Research Quarterly for Exercise and Sport, 69*(2), 111-128. doi:10.1080/02701367.1998.10607677

Williams, A. M., & Reilly, T. (2000). Talent identification and development in soccer. *Journal of Sports Sciences, 18*, 657-667. doi:10.1080/02640410050120041

Williams, H. G., Pfeiffer, K. A., O'Neill, J. R., Dowda, M., McIver, K. L., Brown, W. H., & Pate, R. R. (2008). Motor Skill Performance and Physical Activity in Preschool Children. *Obesity, 16*(6), 1421-1426. doi:10.1038/oby.2008.214

9. Appendix A: Leagues included in *FIFA* per year

Table 6: Leagues included in FIFA *per year*

League	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18	'19	'20
A-League		x	x	x	x	x	x	x	x	x	x	x	x	x
Airtricity League[1]		x	x	x	x	x	x	x	x	x	x	x	x	x
Allsvenskan	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Belgian Pro League	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Bundesliga (AT)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Bundesliga (DE)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Campeonato Nacional[2]								x	x	x	x	x	x	x
Casa Liga 1														x
Championship	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3. Liga												x	x	x
Ekstraklasa	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Elitserien[3]	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Eredivisie	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Gambrinus Liga		x	x	x	x	x	x							
JI League											x	x	x	x
K League Classic	x	x	x	x	x	x	x	x	x	x	x	x	x	x
League One	x	x	x	x	x	x	x	x	x	x	x	x	x	x
League Two	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Liga Dimayor[4]								x	x	x	x	x	x	x
Liga do Brasil[5]	x	x	x	x	x	x	x	x			x	x	x	x
Liga MX	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Liga Portuguesa[6]	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ligue 1	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ligue 2	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Major League Soccer	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Premier League (RU)				x	x	x	x	x	x	x	x	x		
Premier League (UK)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Premiership[7]	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Primera División (AR)								x	x	x	x	x	x	x
Primera División (ES)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Saudi Football League[8]							x	x	x	x	x	x	x	x
Segunda División	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Serie A	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Serie B	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Super League (CH)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Super League (CN)													x	x
Süper Lig	x	x	x	x	x	x	x		x	x	x	x	x	x
Superliga	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2. Bundesliga	x	x	x	x	x	x	x	x	x	x	x	x	x	x

[1] League of Ireland before 2013
[2] Primera División before 2015
[3] Tippeligaen before 2018
[4] Categoría Primera A before 2016
[5] Campeonato Brasileiro before 2014
[6] Primeira Liga before 2014
[7] Premier League before 2013
[8] ALJ League before 2020

10. Appendix B: Descriptive statistics

Table 7: Descriptive statistics of the continuous variables

Variable	Obs.	NA's	Mean	Std. dev.	Min.	Max.
Player attributes						
Age	158,062	0	24.72	4.72	15.00	47.00
Height	158,062	0	181.50	6.59	154.00	208.30
Weight	158,062	0	75.88	6.95	45.80	110.20
Football skills						
Overall	157,867	195	66.34	7.25	33.00	94.00
Attacking: crossing	157,832	230	51.11	17.88	2.00	95.00
Attacking: finishing	157,799	263	46.46	19.16	1.00	97.00
Attacking: heading accuracy	157,833	229	53.96	17.13	1.00	95.00
Attacking: short passing	157,866	196	59.02	14.86	3.00	97.00
Attacking: volleys	156,380	1,682	45.01	17.95	1.00	93.00
Skill: dribbling	157,840	222	55.56	18.40	1.00	97.00
Skill: curve	156,404	1,653	48.59	18.25	2.00	94.00
Skill: free kick accuracy	157,843	219	45.00	17.65	1.00	97.00
Skill: long passing	157,867	195	53.63	15.10	3.00	97.00
Skill: ball control	157,867	195	59.29	16.28	5.00	97.00
Movement: acceleration	157,867	195	65.46	13.88	11.00	97.00
Movement: sprint speed	157,867	195	65.77	13.61	11.00	97.00
Movement: agility	156,434	1,628	63.89	13.87	11.00	96.00
Movement: reactions	157,867	195	62.89	9.45	20.00	96.00
Movement: balance	156,434	1,628	64.23	13.45	10.00	99.00
Power: shot power	157,866	196	57.45	16.64	2.00	96.00
Power: jumping	156,434	1,628	65.61	11.43	13.00	97.00
Power: stamina	157,867	195	64.33	14.75	10.00	97.00
Power: strength	157,867	195	65.88	12.39	12.00	98.00
Power: long shots	157,824	238	48.81	18.81	1.00	96.00
Mentality: aggression	157,867	195	57.44	16.93	2.00	97.00
Mentality: interceptions	157,844	218	48.90	19.86	1.00	96.00
Mentality: positioning	157,755	307	51.69	18.94	2.00	96.00
Mentality: vision	156,392	1,670	54.48	14.89	1.00	97.00
Mentality: penalties	157,862	200	51.04	15.88	2.00	96.00
Defending: marking	157,816	246	45.79	20.81	1.00	96.00
Defending: standing tackle	157,839	223	48.28	21.39	1.00	95.00
Defending: sliding tackle	156,402	1,660	46.28	21.28	2.00	95.00
Goalkeeping: diving	153,278	4,784	16.32	17.99	1.00	94.00
Goalkeeping: handling	153,361	4,701	17.32	16.90	1.00	93.00
Goalkeeping: kicking	153,437	4,625	20.90	20.82	1.00	97.00
Goalkeeping: positioning	153,357	4,705	17.38	17.09	1.00	96.00
Goalkeeping: reflexes	153,343	4,719	17.76	18.17	1.00	96.00

Table 8: Descriptive statistics of the categorical variables

Variable	Category	Frequency	Relative frequency
Player attributes			
Preferred position	Goalkeeper (GK)	16,568	0.105
	Centre back (CB)	27,283	0.173
	Left back (LB)	11,850	0.075
	Right back (RB)	11,807	0.075
	Centre defensive midfielder (CDM)	13,085	0.083
	Centre midfielder (CM)	17,854	0.113
	Centre attacking midfielder (CAM)	9,563	0.061
	Left midfielder (LM)	8,628	0.055
	Right midfielder (RM)	8,293	0.052
	Striker (ST)	24,494	0.155
	Left winger (LW)	3,282	0.055
	Right winger (RW)	3,552	0.022
	NA	1,803	0.011
Preferred foot	Right	120,448	0.762
	Left	37,419	0.237
	NA	195	0.001
Football skills			
Attacking work rate	Low	8,147	0.052
	Medium	112,148	0.710
	High	35,764	0.226
	NA	2,003	0.013
Defensive work rate	Low	18,112	0.115
	Medium	116,562	0.737
	High	23,193	0.147
	NA	195	0.001

11. Appendix C: Scatterplots of the fit per model

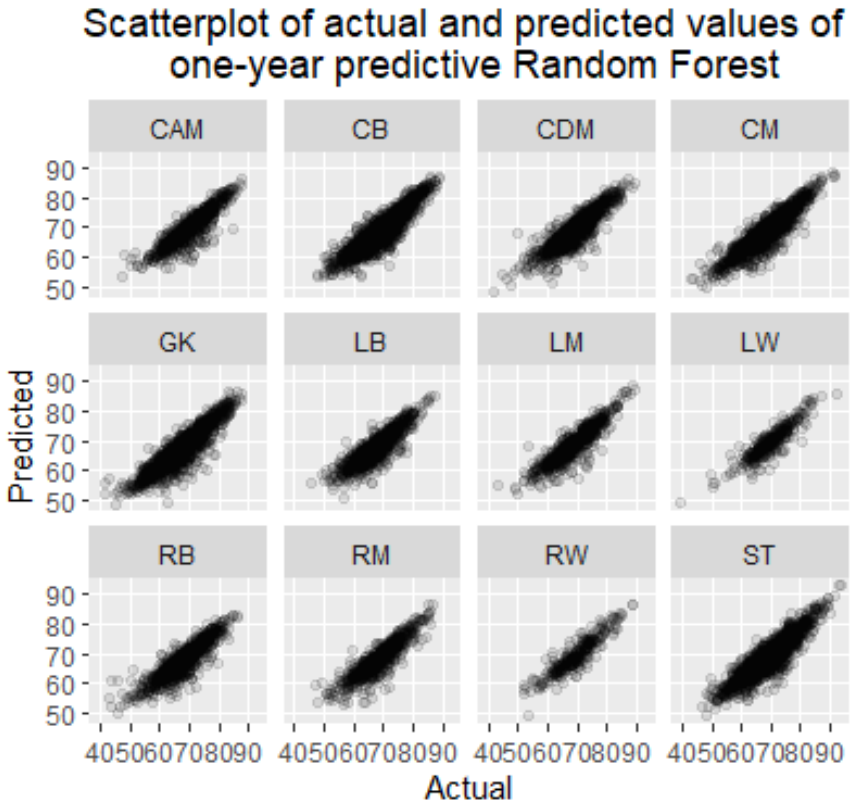


Figure 16: Scatterplot of actual and predicted values per position of the one-year predictive Random Forest

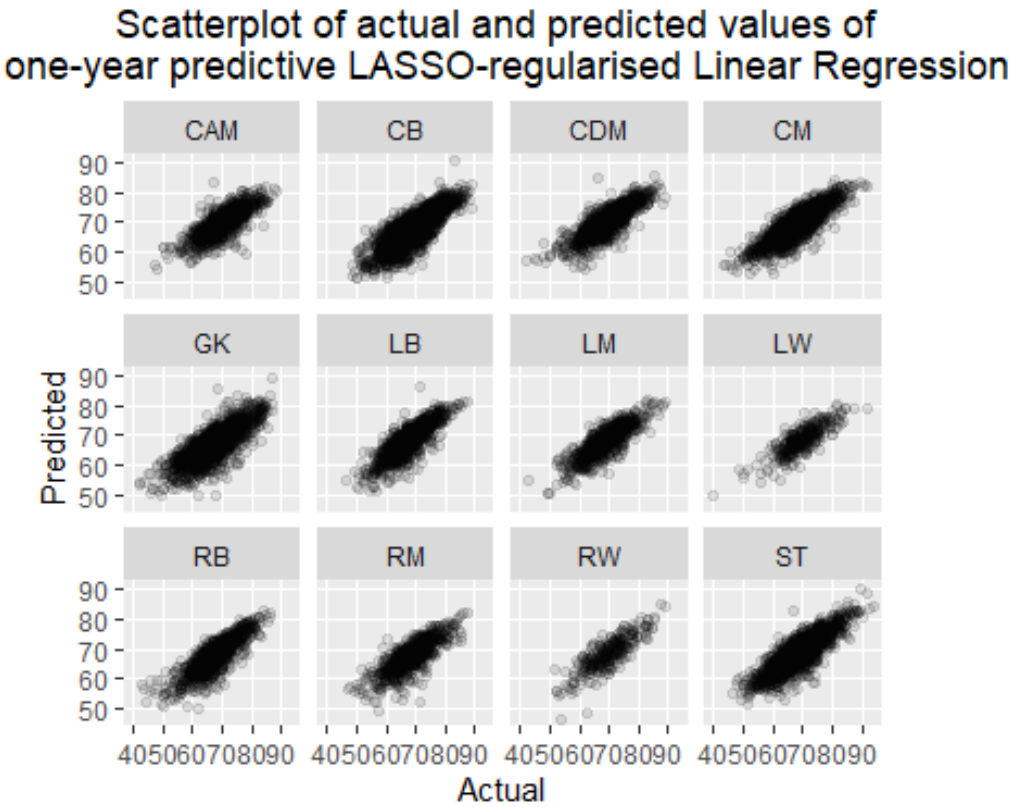


Figure 17: Scatterplot of actual and predicted values per position of the one-year predictive LASSO-regularised Linear Regression

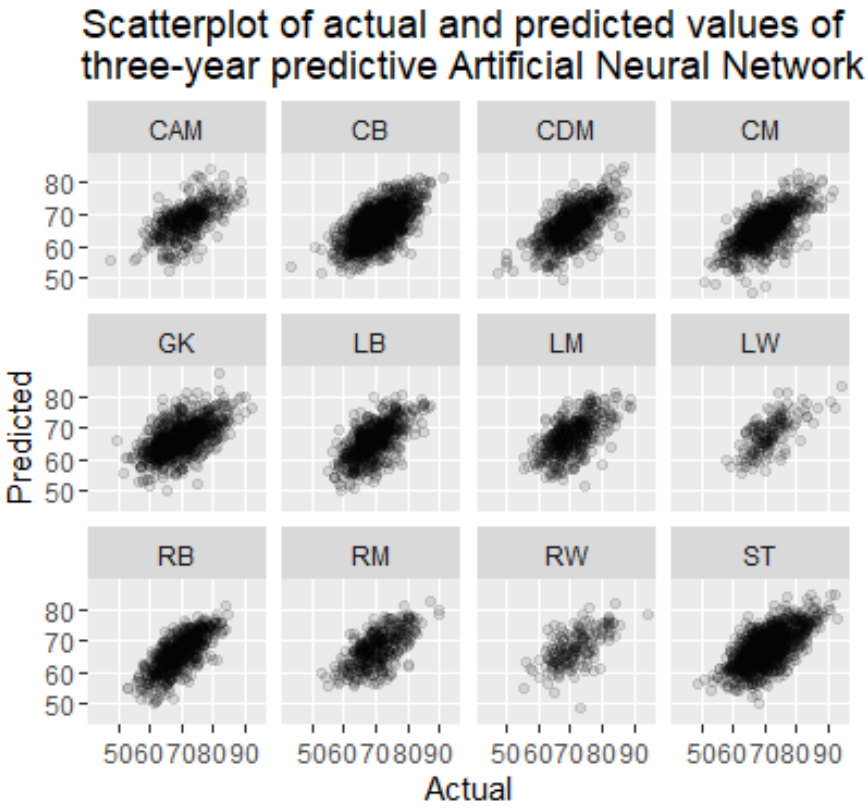


Figure 18: Scatterplot of actual and predicted values per position of the three-year predictive Artificial Neural Network

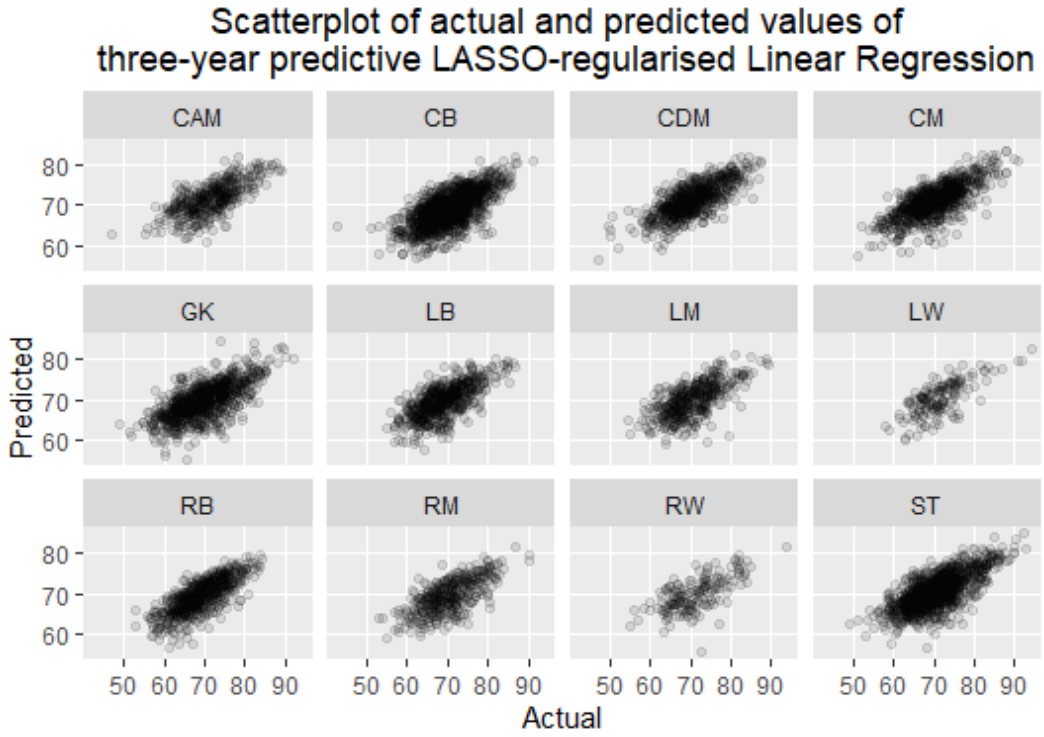
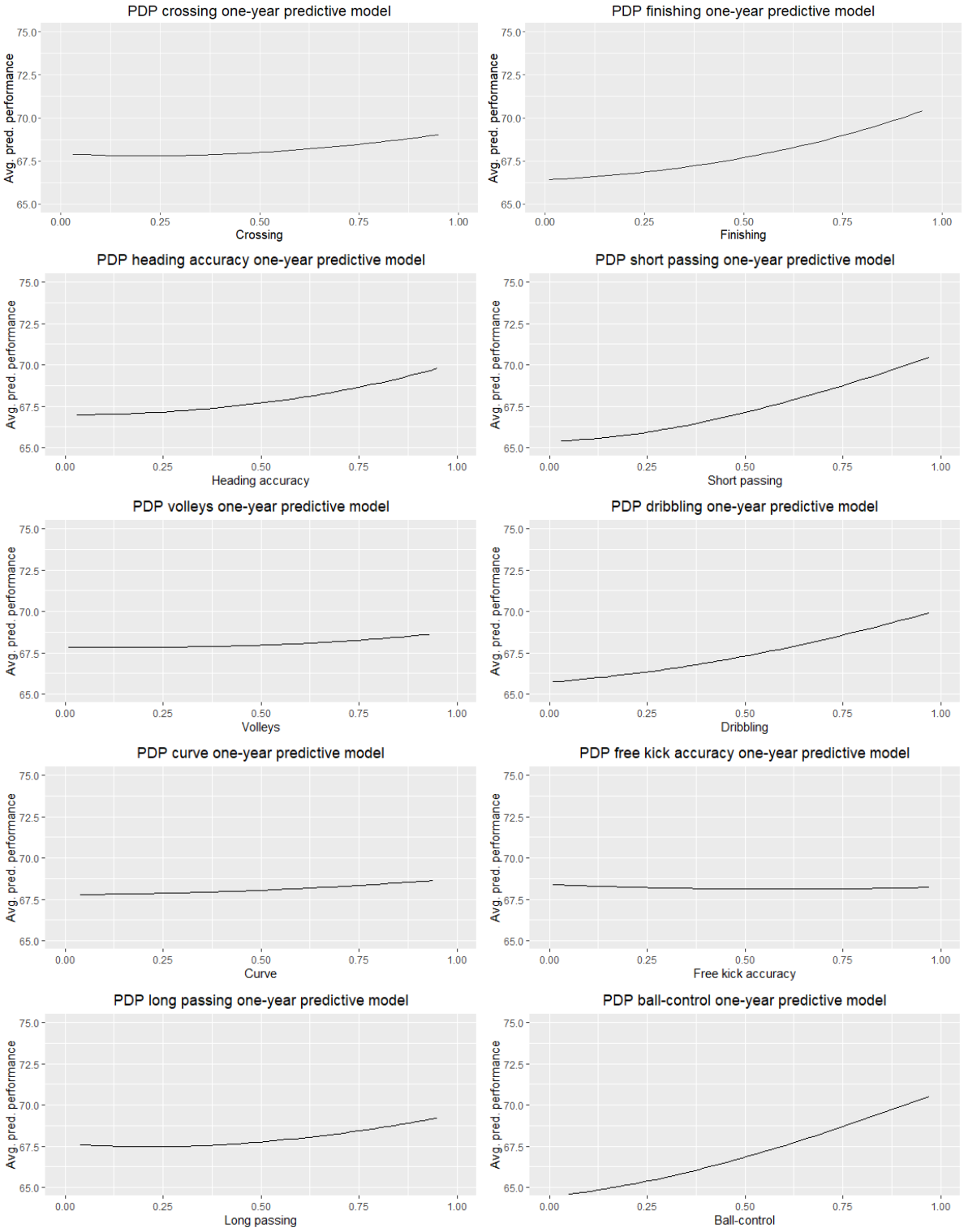
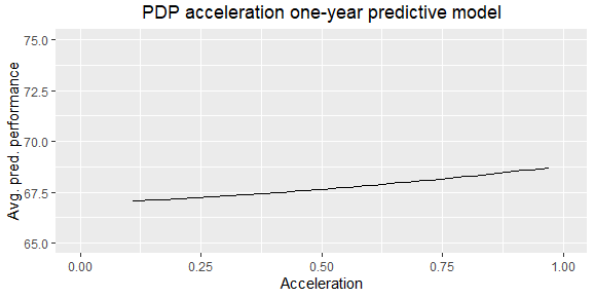
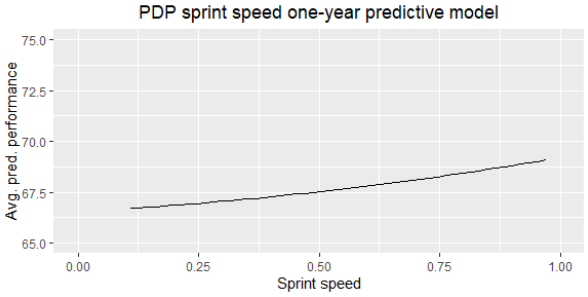


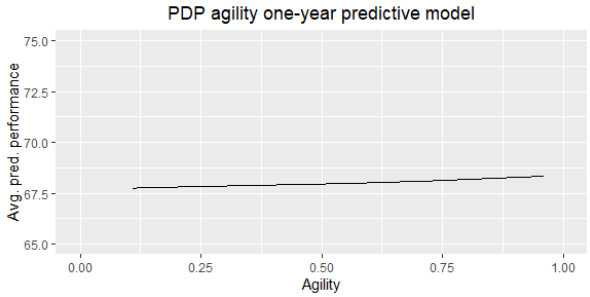
Figure 19: Scatterplot of actual and predicted values per position of the three-year predictive Artificial Neural Network

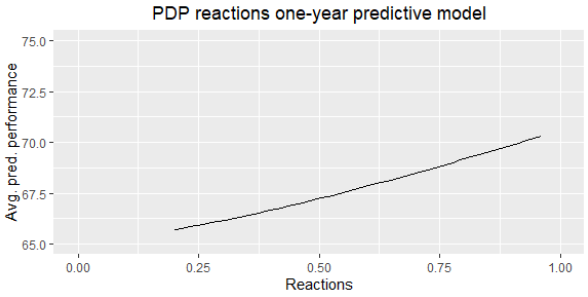
12. Appendix D: Partial Dependence Plots of main effect in one-year predictive Artificial Neural Network

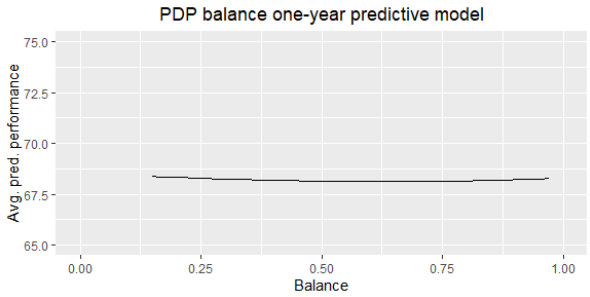


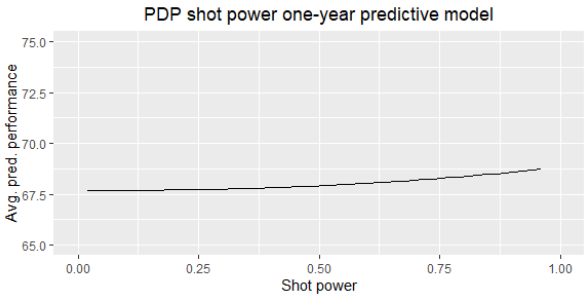

PDP acceleration one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Acceleration

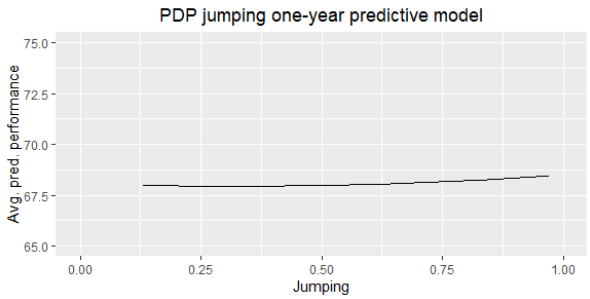

PDP sprint speed one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Sprint speed

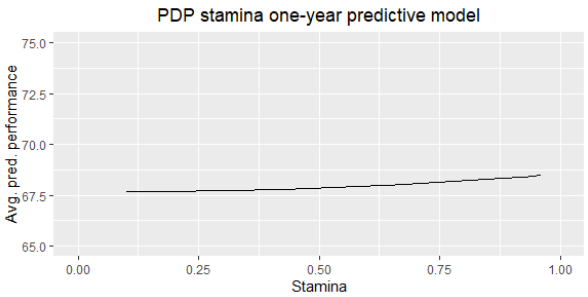

PDP agility one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Agility

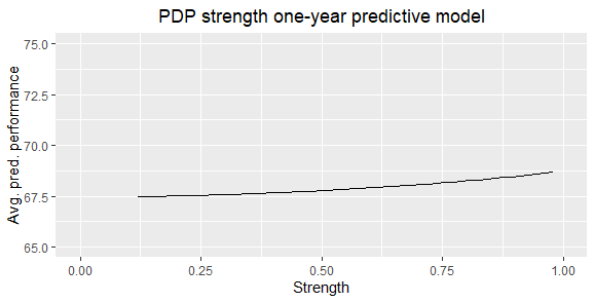

PDP reactions one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Reactions

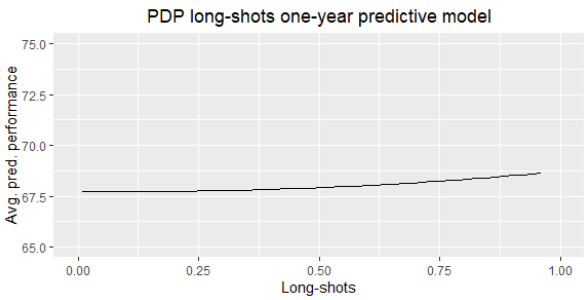

PDP balance one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Balance

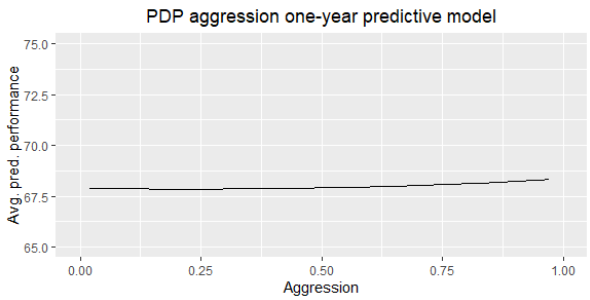

PDP shot power one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Shot power

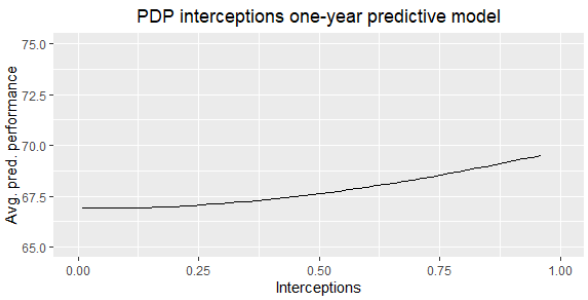

PDP jumping one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Jumping

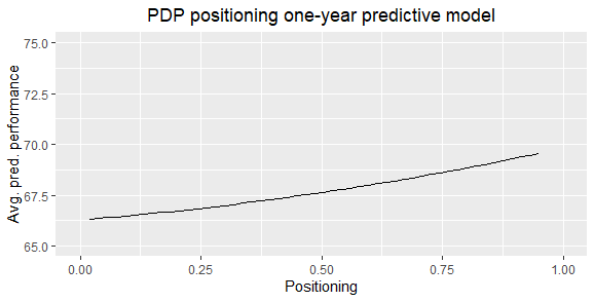

PDP stamina one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Stamina

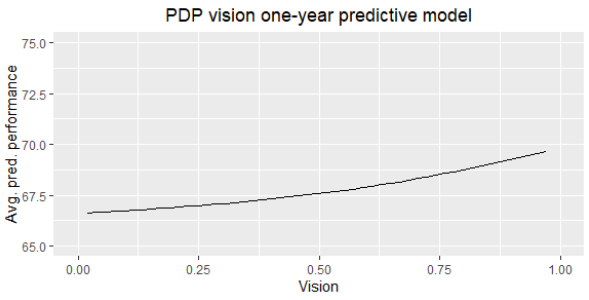

PDP strength one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Strength

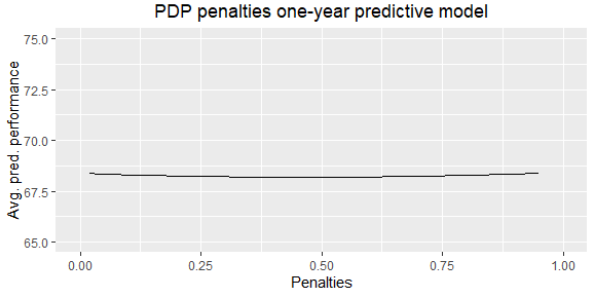

PDP long-shots one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Long-shots

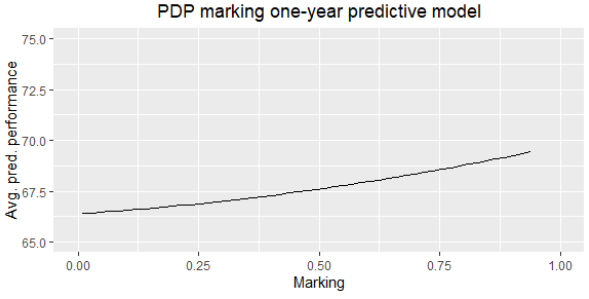

PDP aggression one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Aggression

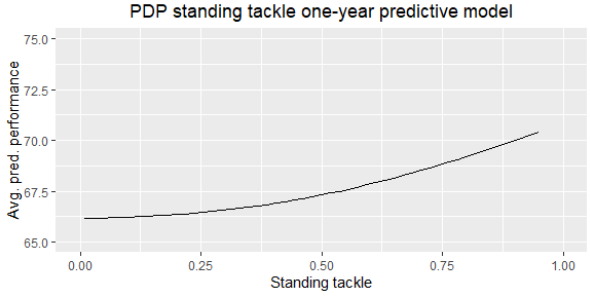

PDP interceptions one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Interceptions

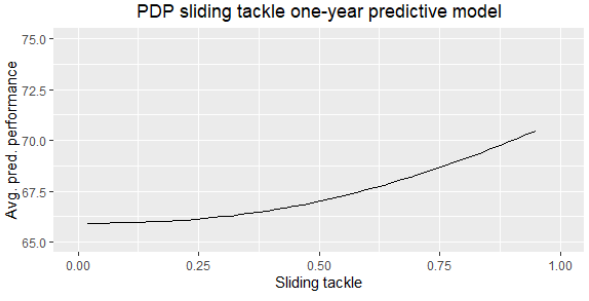

PDP positioning one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Positioning

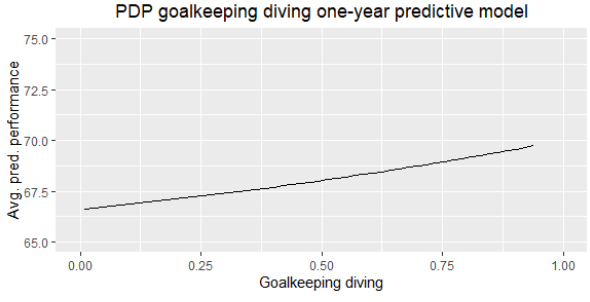

PDP vision one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Vision

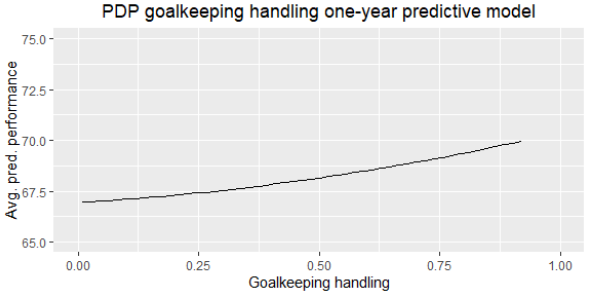

PDP penalties one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Penalties

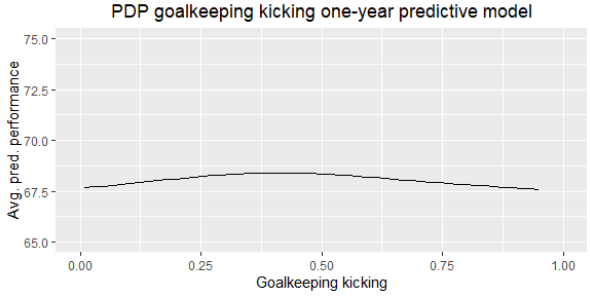

PDP marking one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Marking

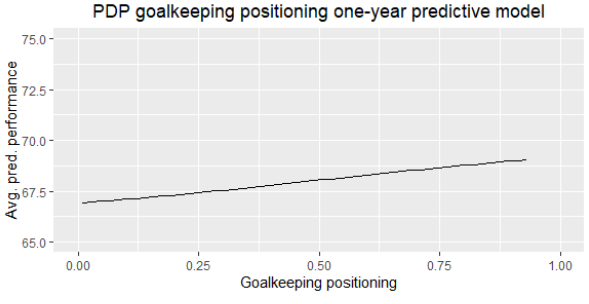

PDP standing tackle one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Standing tackle

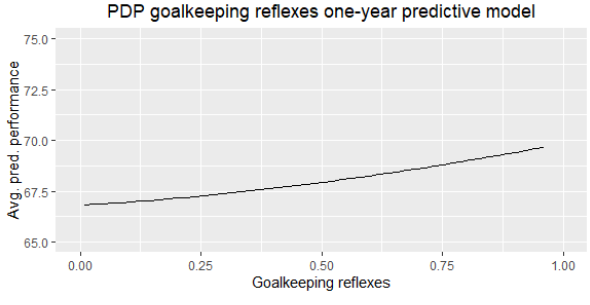

PDP sliding tackle one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Sliding tackle

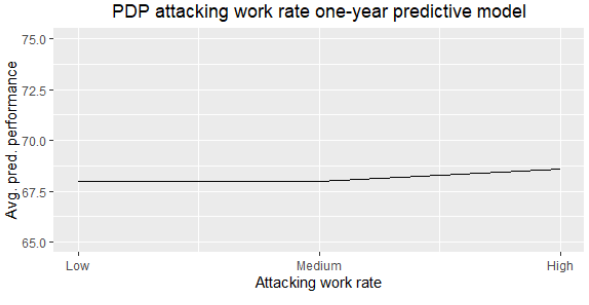

PDP goalkeeping diving one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Goalkeeping diving


PDP goalkeeping handling one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Goalkeeping handling


PDP goalkeeping kicking one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Goalkeeping kicking


PDP goalkeeping positioning one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Goalkeeping positioning


PDP goalkeeping reflexes one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Goalkeeping reflexes


PDP attacking work rate one-year predictive model
Avg. pred. performance
75.0
72.5
70.0
67.5
65.0
Low
Medium
High
Attacking work rate

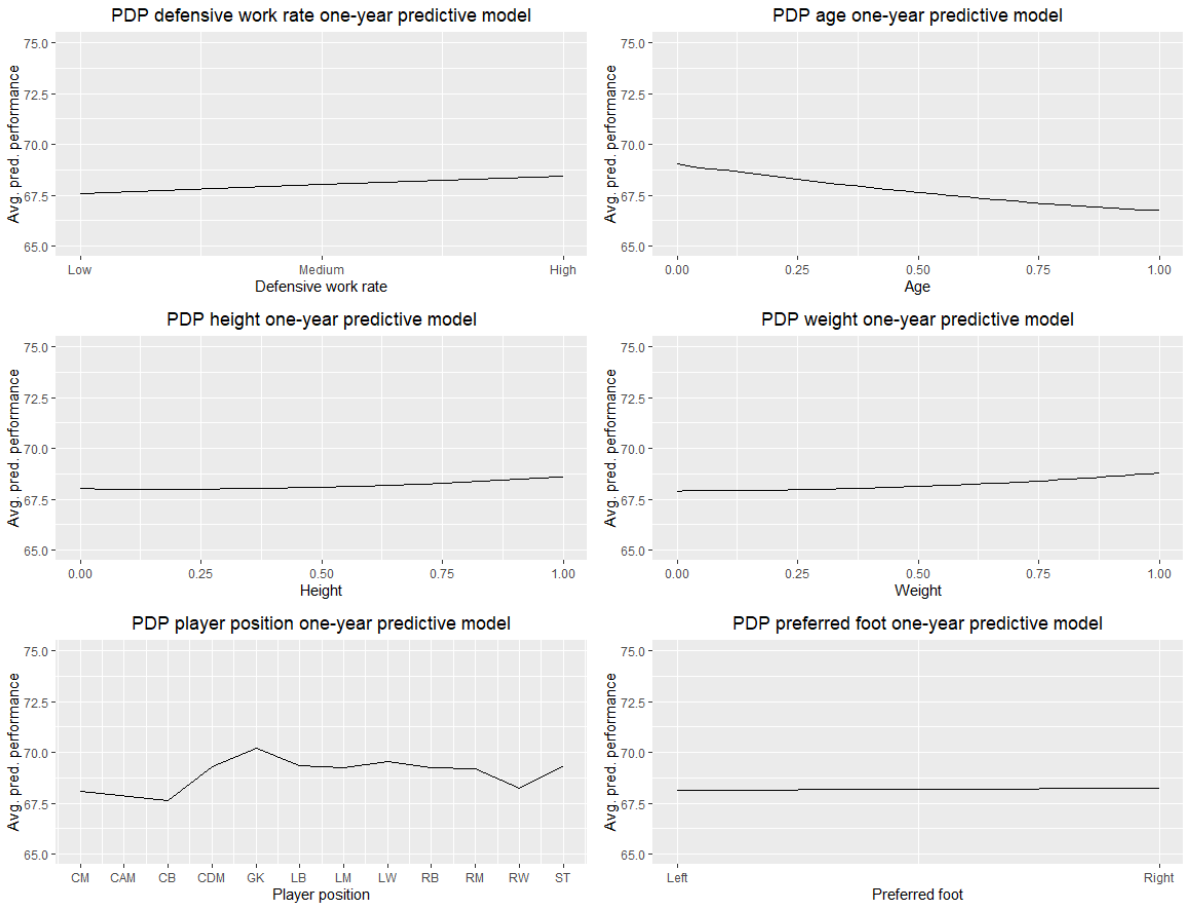
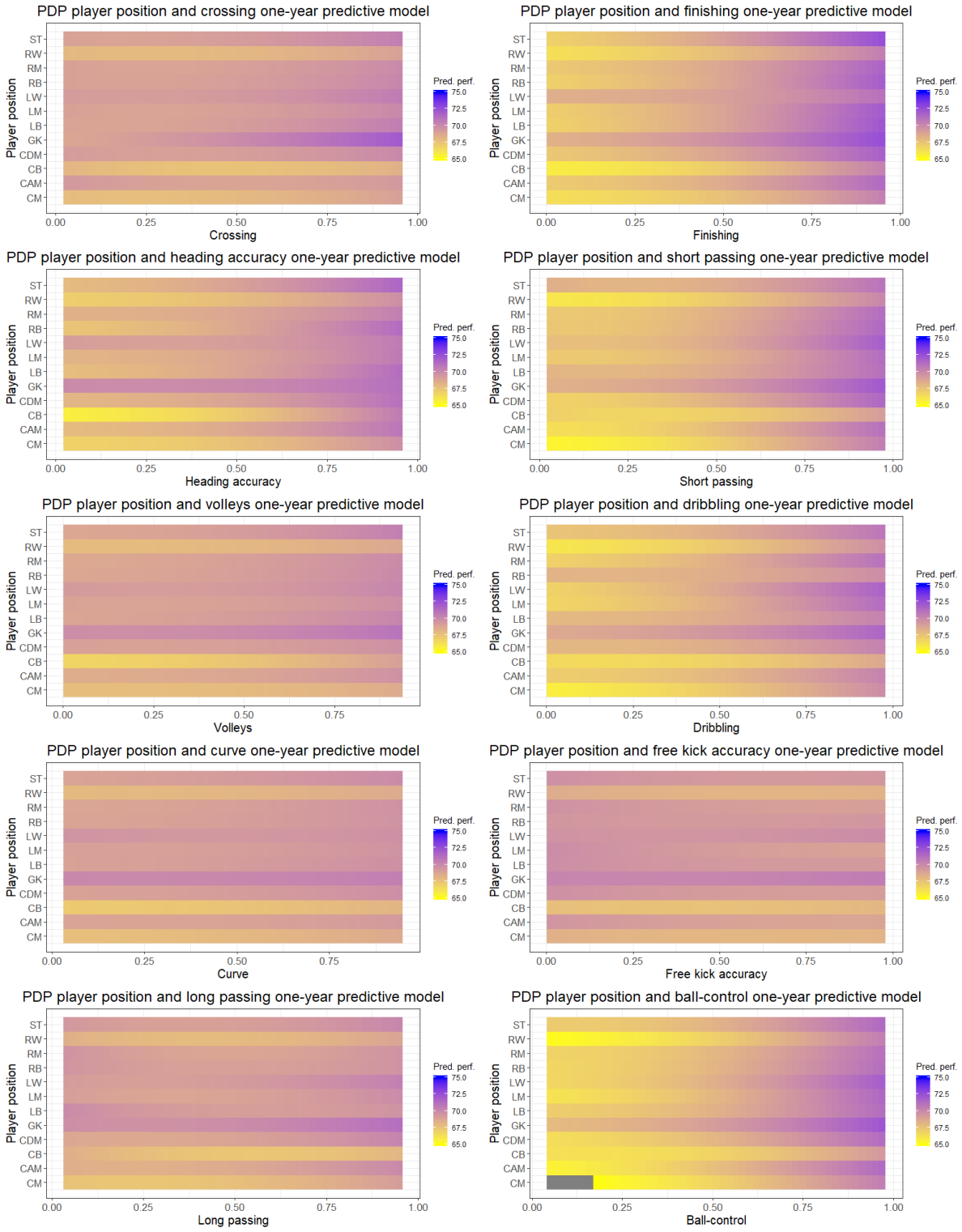
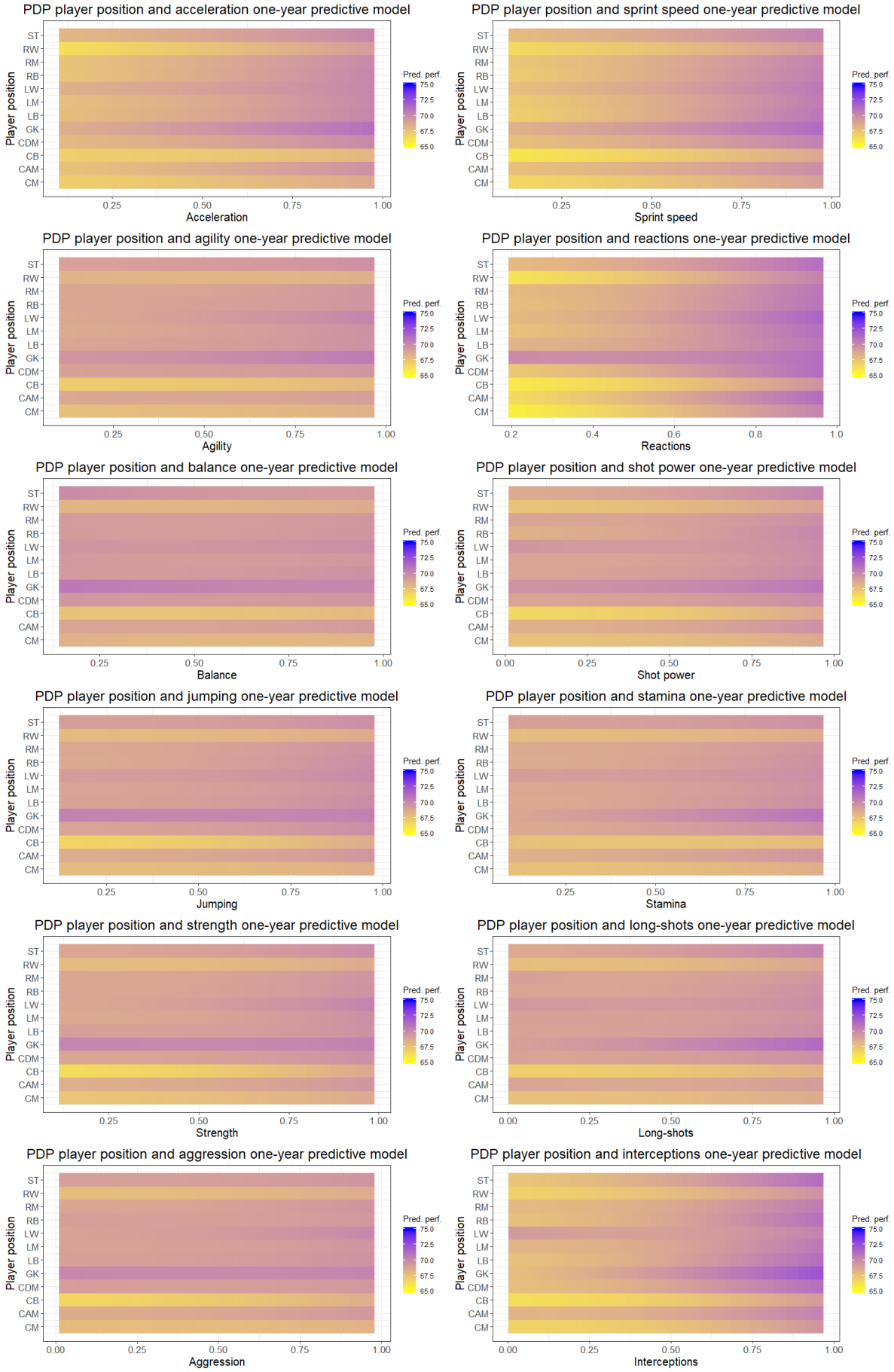
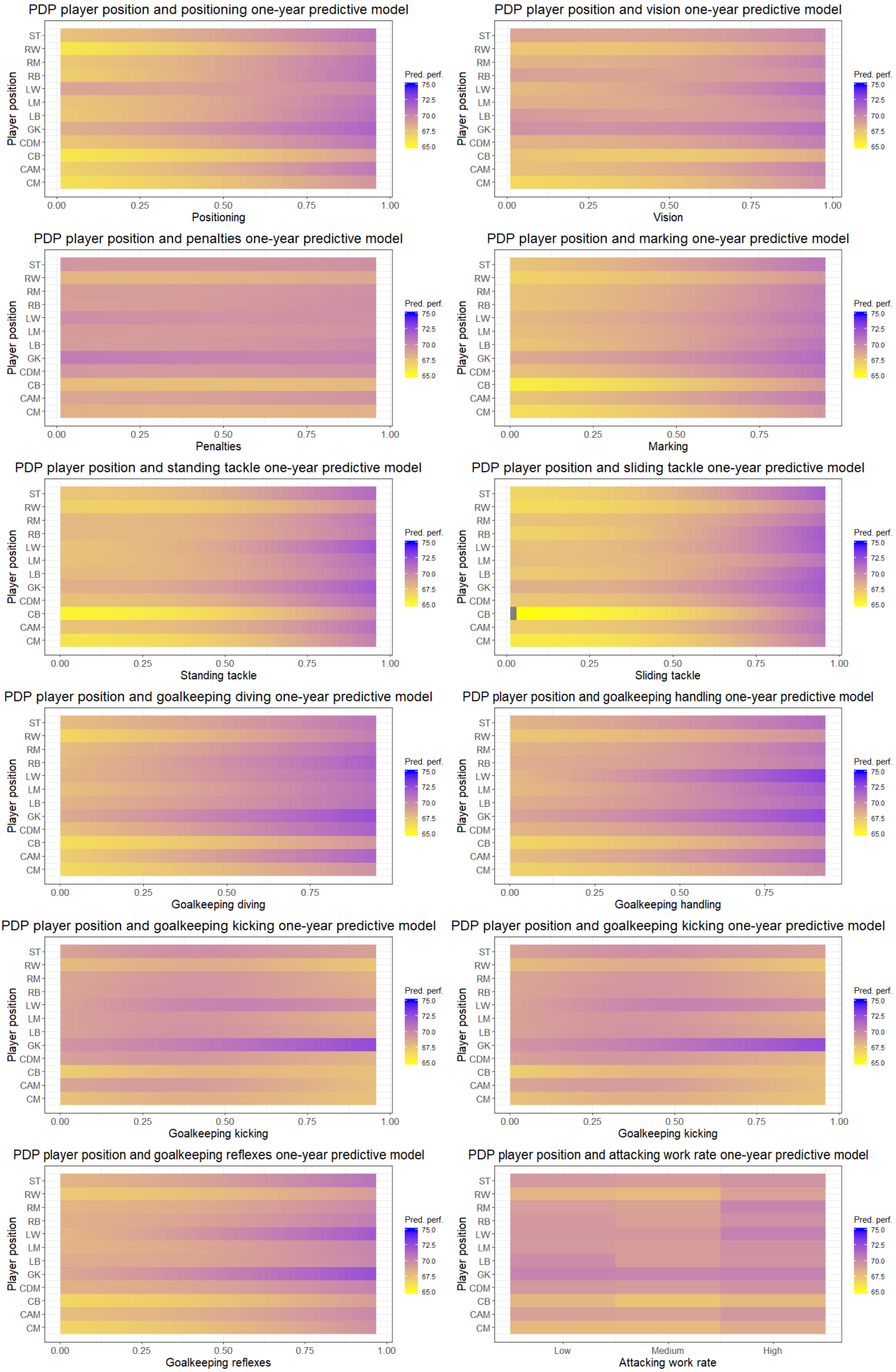


Figure 20: Partial Dependence Plots of all variables in the one-year predictive Artificial Neural Network

13. Appendix E: Partial Dependence Plots of interaction effect of player position in one-year predictive Artificial Neural Network




PDP player position and acceleration one-year predictive model
PDP player position and sprint speed one-year predictive model
PDP player position and agility one-year predictive model
PDP player position and reactions one-year predictive model
PDP player position and balance one-year predictive model
PDP player position and shot power one-year predictive model
PDP player position and jumping one-year predictive model
PDP player position and stamina one-year predictive model
PDP player position and strength one-year predictive model
PDP player position and long-shots one-year predictive model
PDP player position and aggression one-year predictive model
PDP player position and interceptions one-year predictive model
Player position
ST
RW
RM
RB
LW
LM
LB
GK
CDM
CB
CAM
CM
Pred. perf.
75.0
72.5
70.0
67.5
65.0
Acceleration
Sprint speed
Agility
Reactions
Balance
Shot power
Jumping
Stamina
Strength
Long-shots
Aggression
Interceptions


PDP player position and positioning one-year predictive model
PDP player position and vision one-year predictive model
PDP player position and penalties one-year predictive model
PDP player position and marking one-year predictive model
PDP player position and standing tackle one-year predictive model
PDP player position and sliding tackle one-year predictive model
PDP player position and goalkeeping diving one-year predictive model
PDP player position and goalkeeping handling one-year predictive model
PDP player position and goalkeeping kicking one-year predictive model
PDP player position and goalkeeping kicking one-year predictive model
PDP player position and goalkeeping reflexes one-year predictive model
PDP player position and attacking work rate one-year predictive model
Player position
ST
RW
RM
RB
LW
LM
LB
GK
CDM
CB
CAM
CM
Pred. perf.
75.0
72.5
70.0
67.5
65.0
0.00
0.25
0.50
0.75
1.00
Positioning
Vision
Penalties
Marking
Standing tackle
Sliding tackle
Goalkeeping diving
Goalkeeping handling
Goalkeeping kicking
Goalkeeping kicking
Goalkeeping reflexes
Low
Medium
High
Attacking work rate

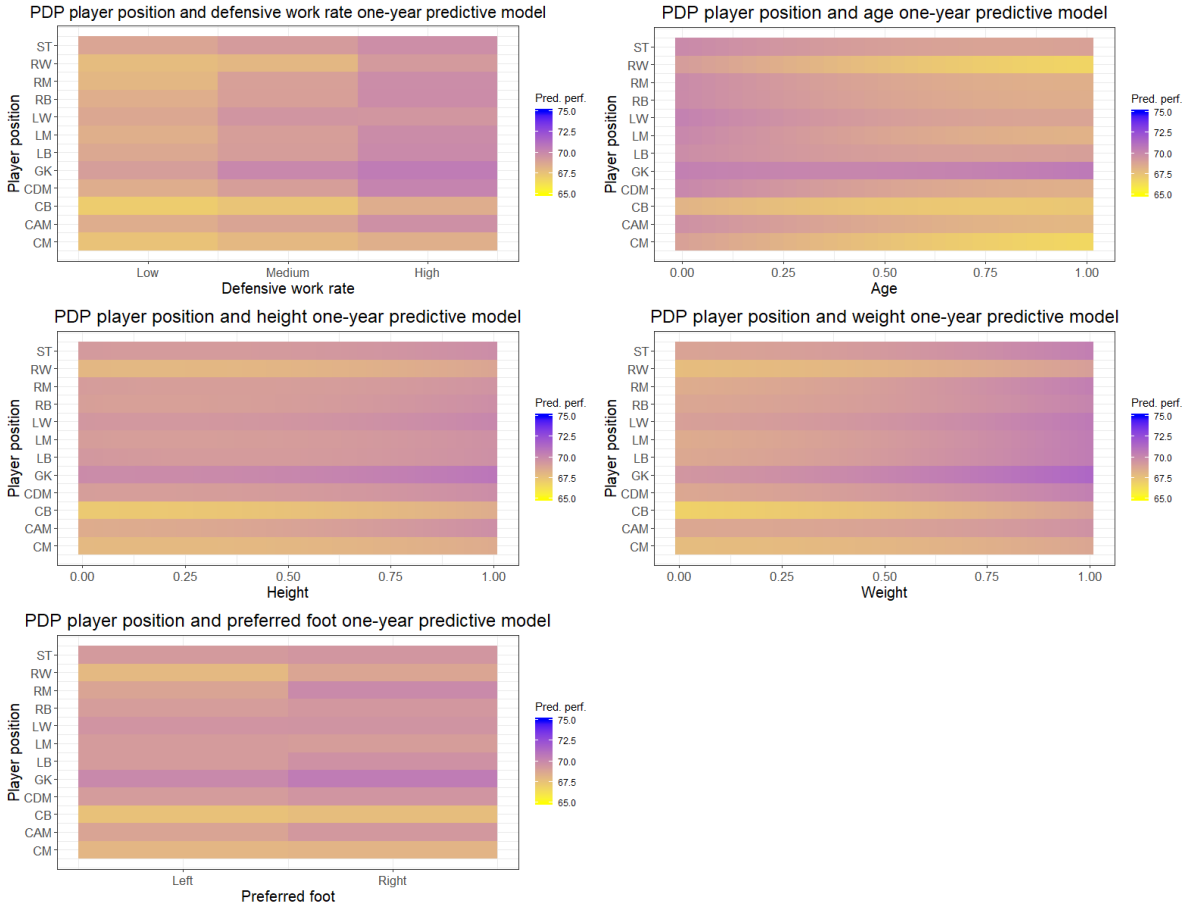
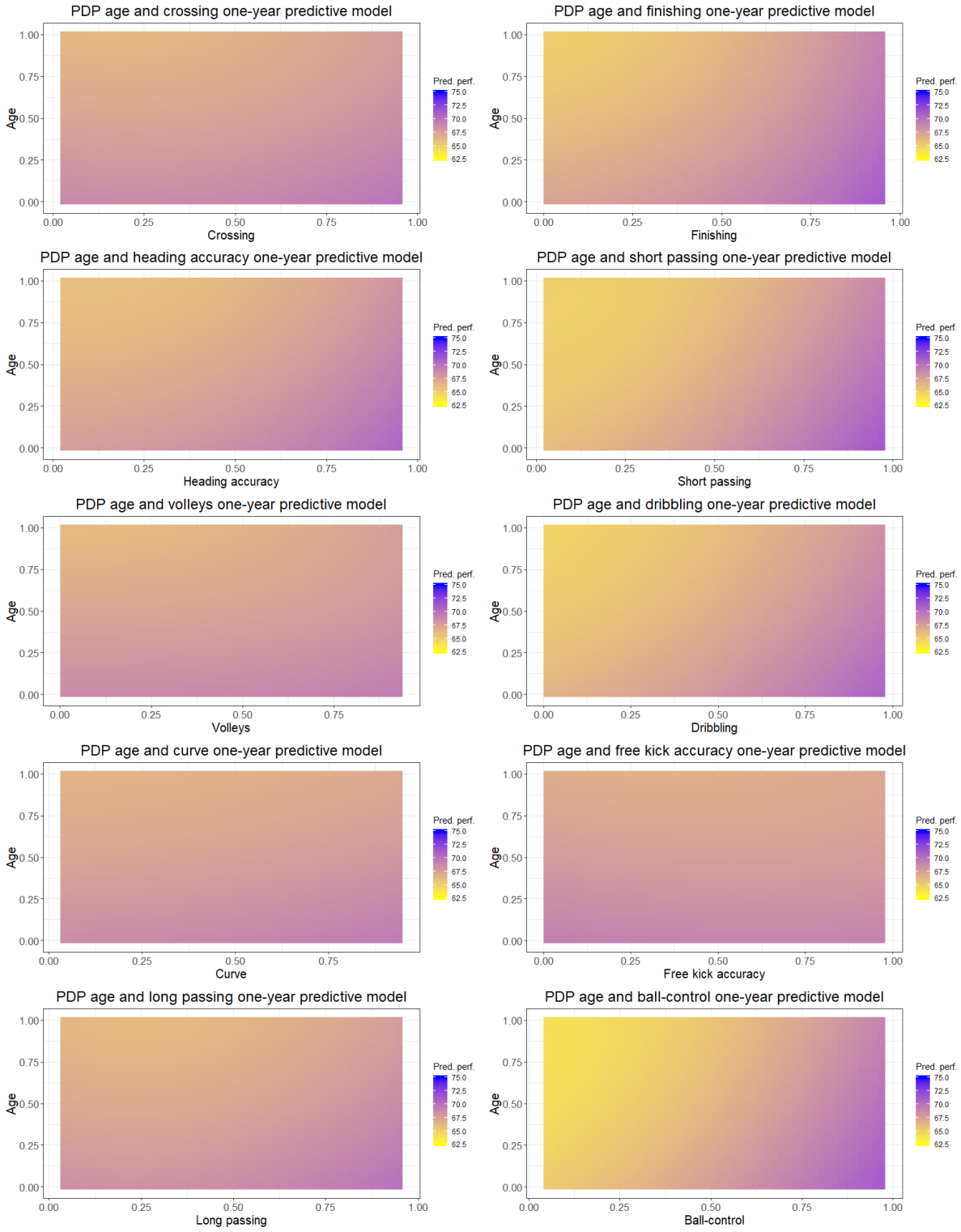
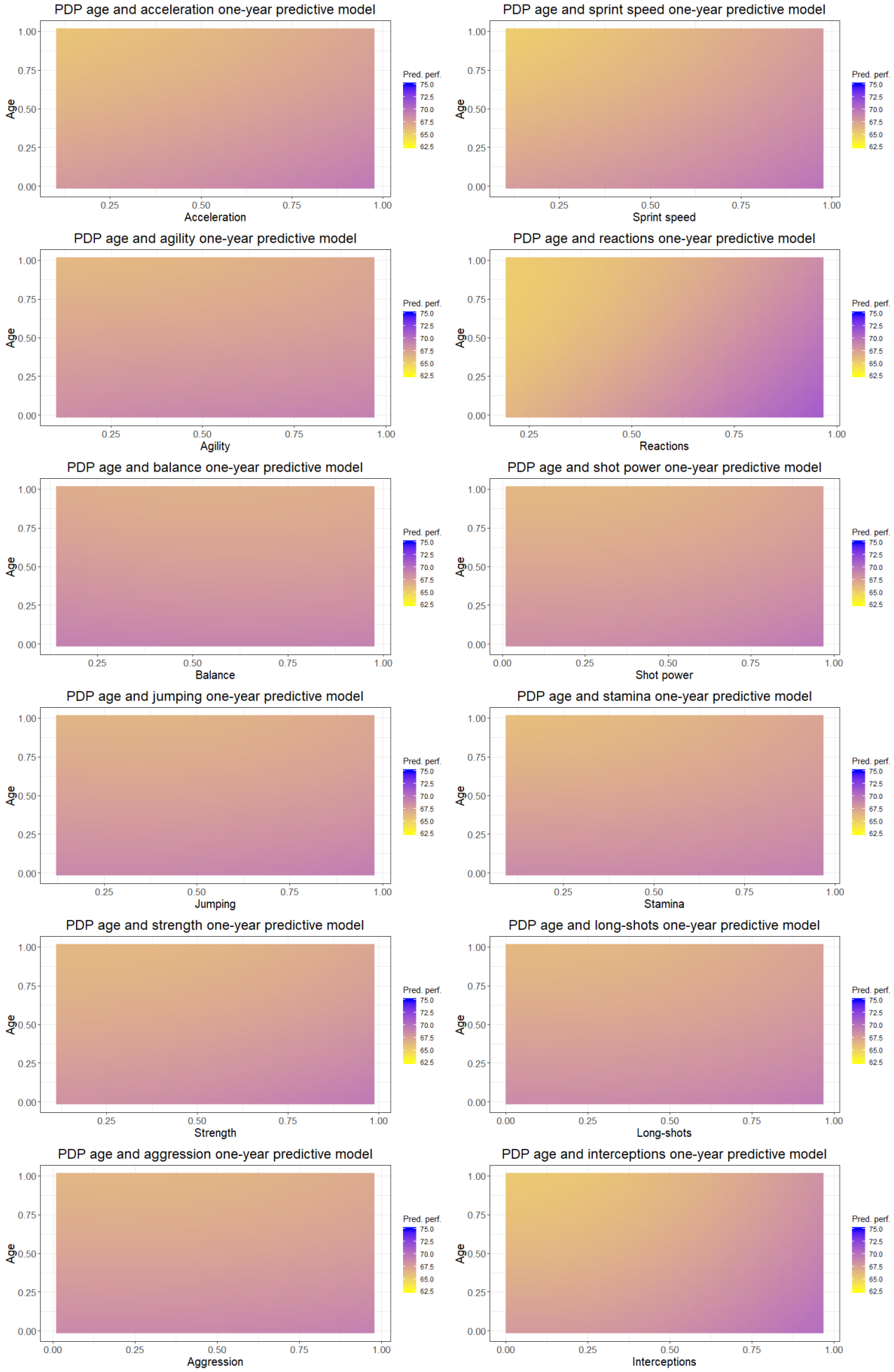
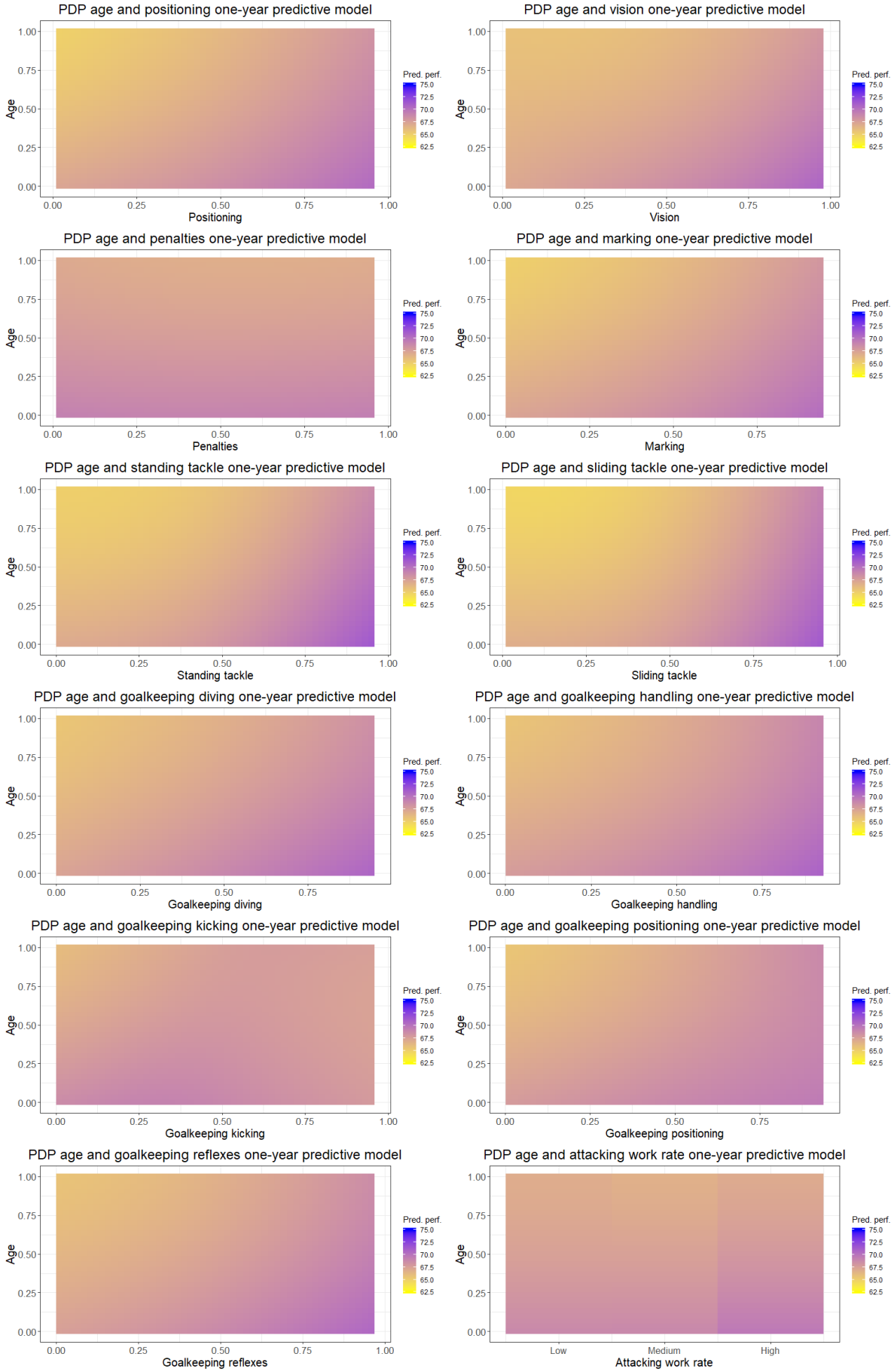


Figure 21: PDPs of all variables in the one-year Artificial Neural Network with player position included as interaction effect

14. Appendix F: Partial Dependence Plots of interaction effect of age in one-year predictive Artificial Neural Network




PDP age and acceleration one-year predictive model
PDP age and sprint speed one-year predictive model
PDP age and agility one-year predictive model
PDP age and reactions one-year predictive model
PDP age and balance one-year predictive model
PDP age and shot power one-year predictive model
PDP age and jumping one-year predictive model
PDP age and stamina one-year predictive model
PDP age and strength one-year predictive model
PDP age and long-shots one-year predictive model
PDP age and aggression one-year predictive model
PDP age and interceptions one-year predictive model
Age
Acceleration
Sprint speed
Agility
Reactions
Balance
Shot power
Jumping
Stamina
Strength
Long-shots
Aggression
Interceptions
Pred. perf.
75.0
72.5
70.0
67.5
65.0
62.5


PDP age and positioning one-year predictive model
PDP age and vision one-year predictive model
PDP age and penalties one-year predictive model
PDP age and marking one-year predictive model
PDP age and standing tackle one-year predictive model
PDP age and sliding tackle one-year predictive model
PDP age and goalkeeping diving one-year predictive model
PDP age and goalkeeping handling one-year predictive model
PDP age and goalkeeping kicking one-year predictive model
PDP age and goalkeeping positioning one-year predictive model
PDP age and goalkeeping reflexes one-year predictive model
PDP age and attacking work rate one-year predictive model
Age
Pred. perf.
75.0
72.5
70.0
67.5
65.0
62.5
Positioning
Vision
Penalties
Marking
Standing tackle
Sliding tackle
Goalkeeping diving
Goalkeeping handling
Goalkeeping kicking
Goalkeeping positioning
Goalkeeping reflexes
Low
Medium
High
Attacking work rate

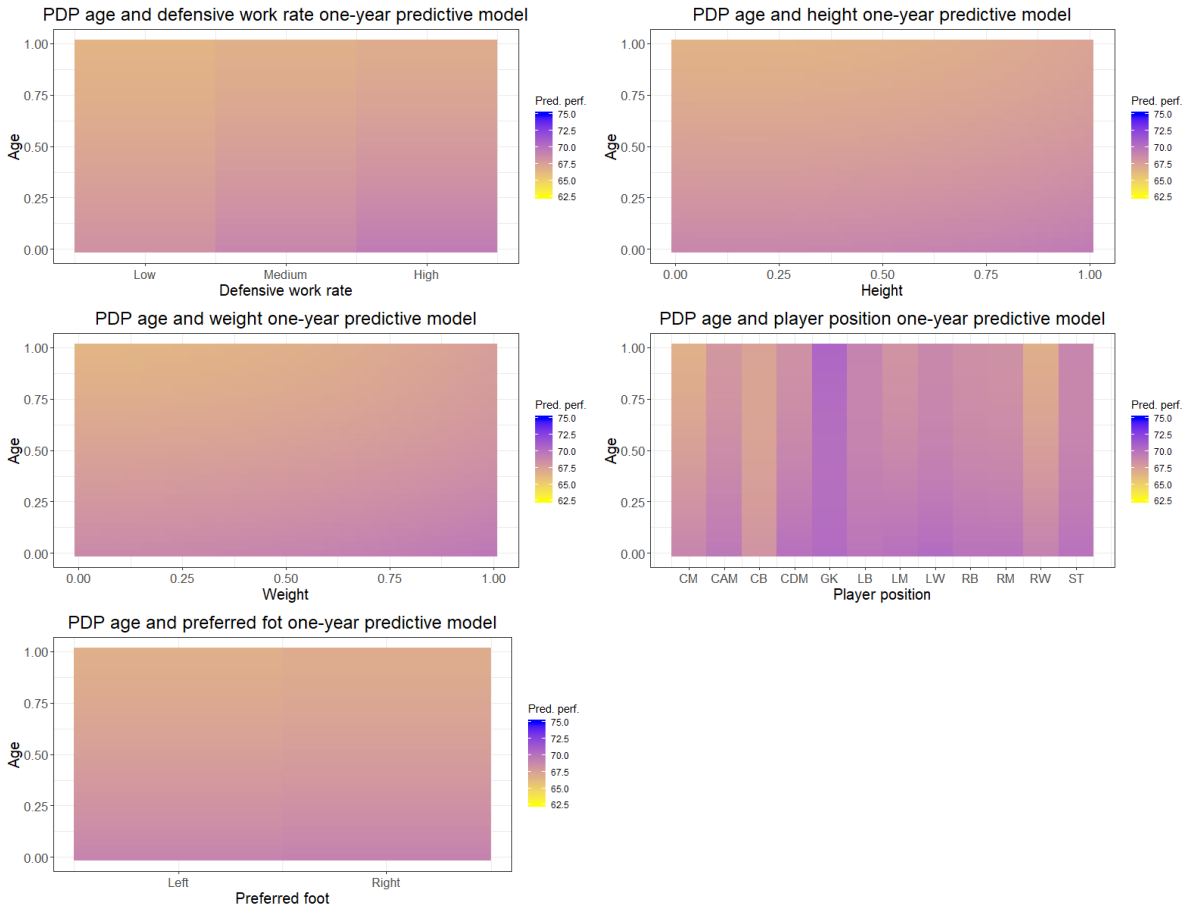
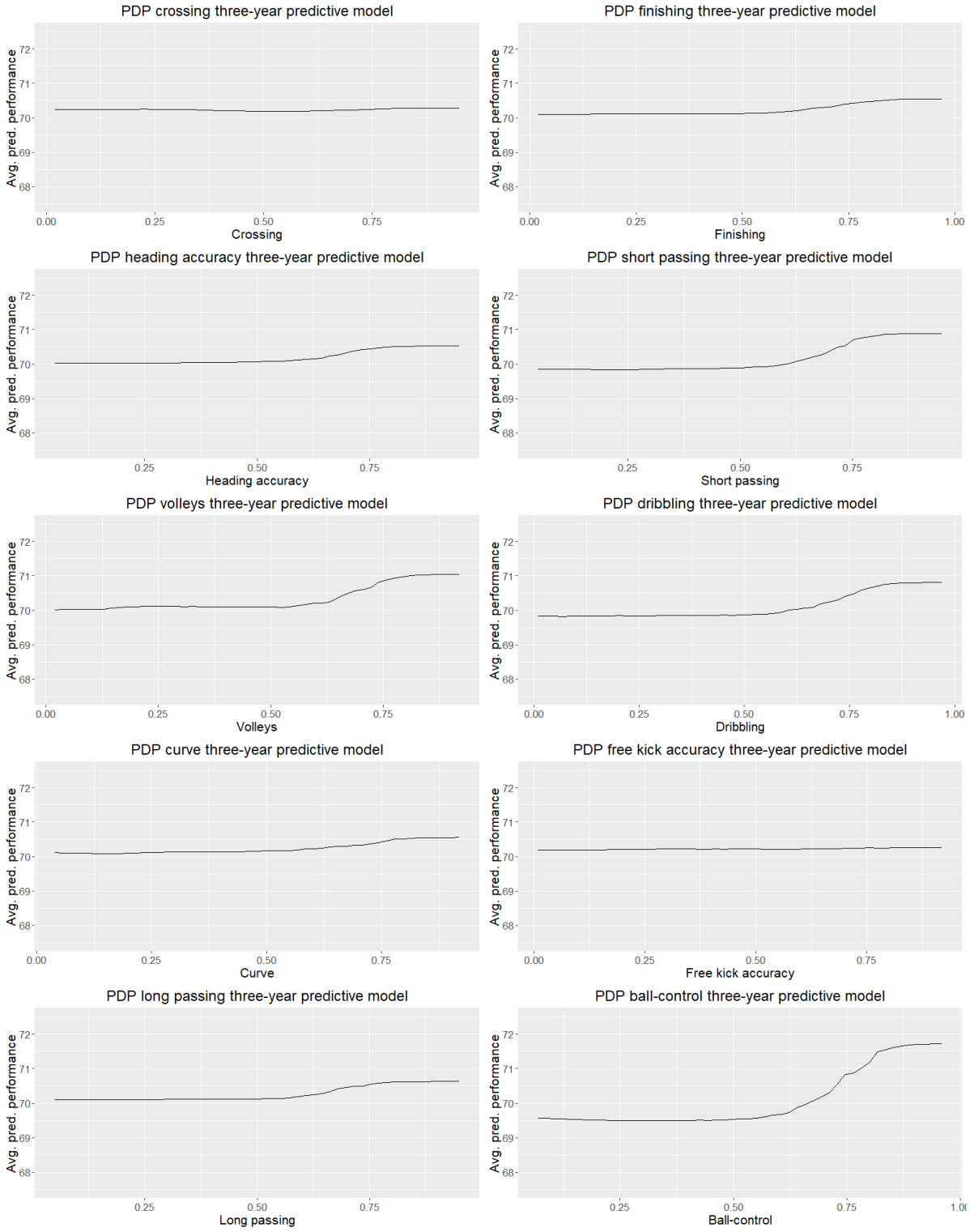
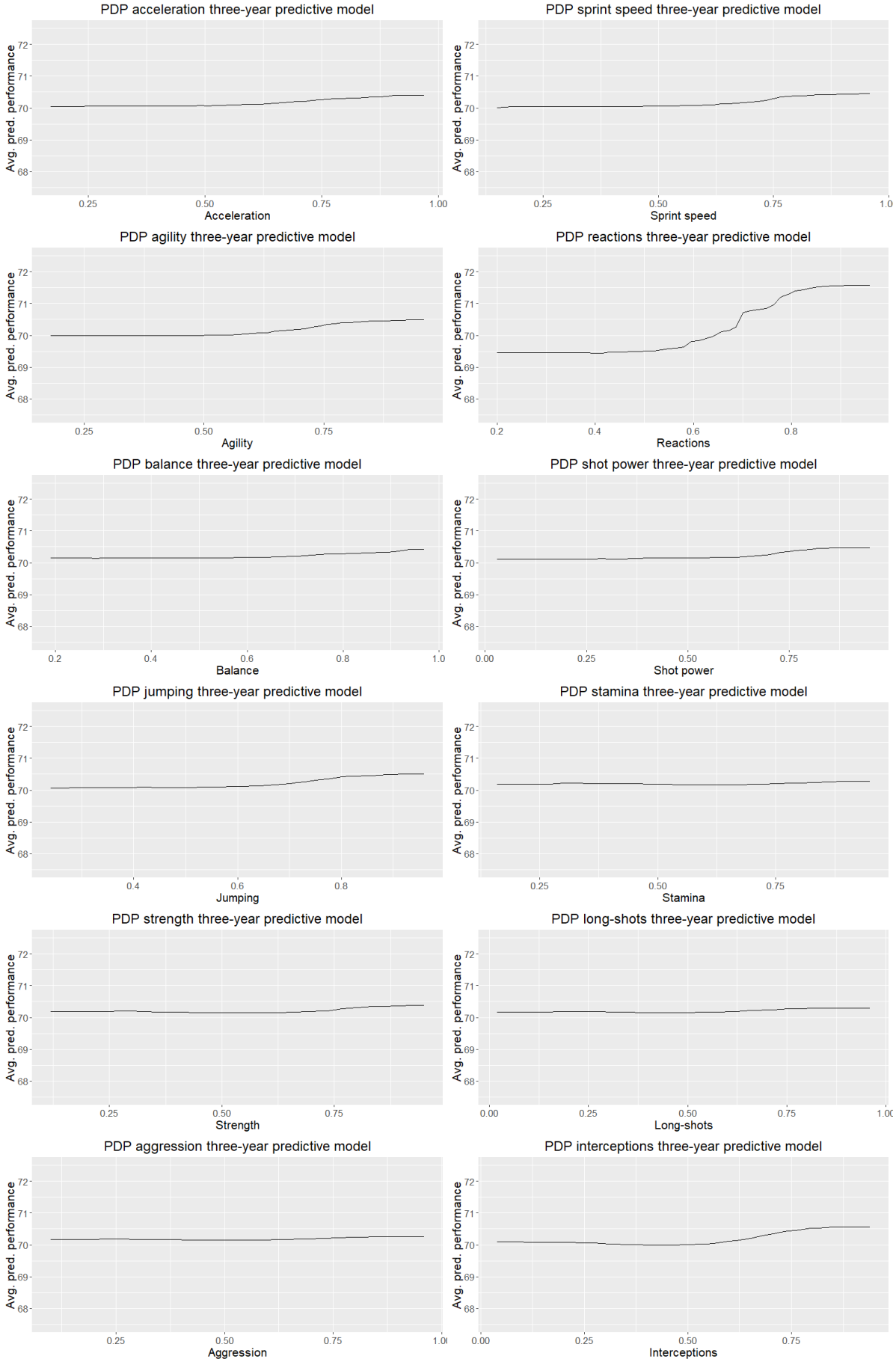
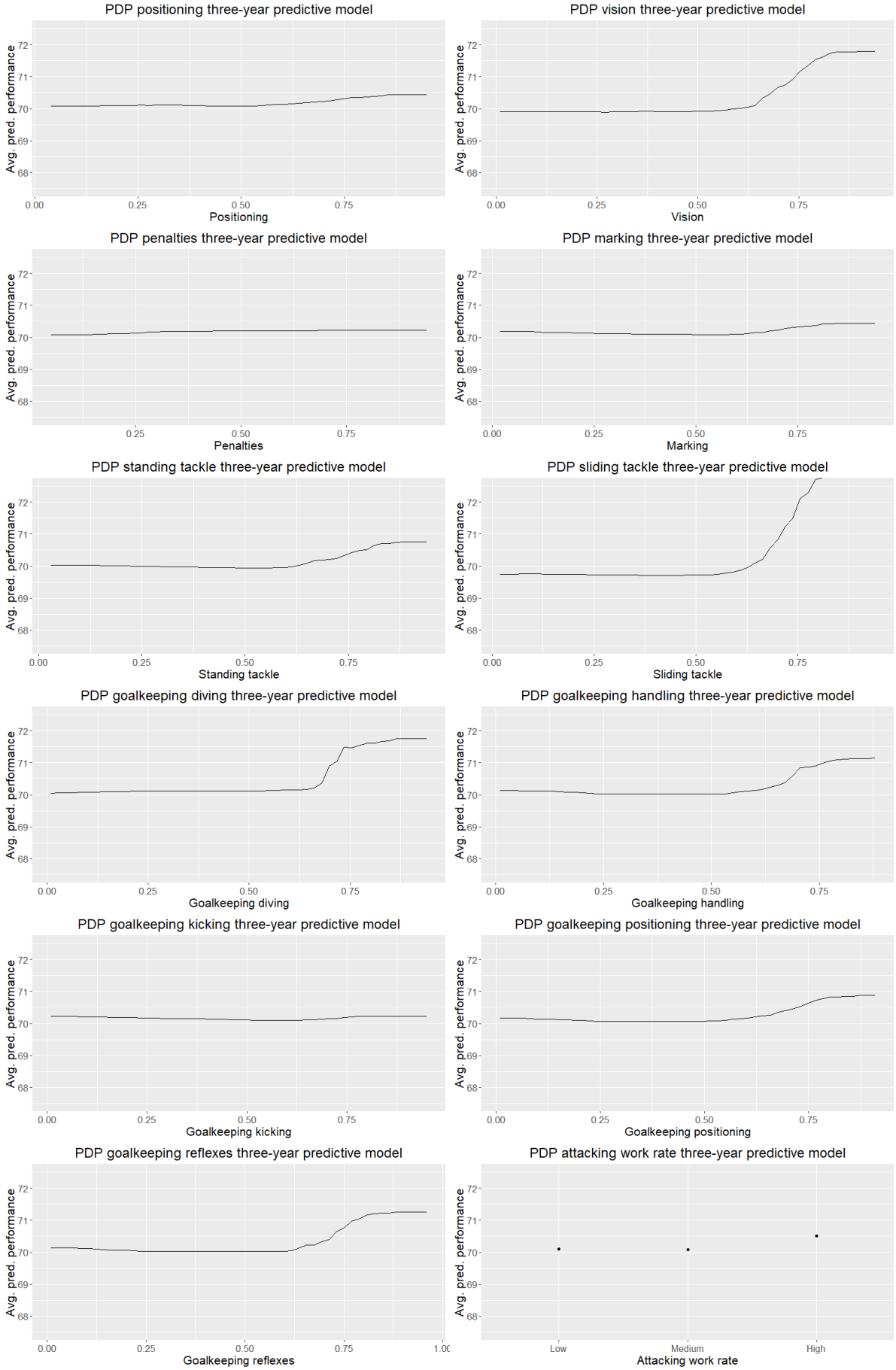


Figure 22: PDPs of all variables in the one-year predictive Artificial Neural Network with age included as interaction effect

15. Appendix G: Partial Dependence Plots of main effect in three-year predictive Random Forest




PDP acceleration three-year predictive model
Avg. pred. performance
Acceleration
PDP sprint speed three-year predictive model
Sprint speed
PDP agility three-year predictive model
Agility
PDP reactions three-year predictive model
Reactions
PDP balance three-year predictive model
Balance
PDP shot power three-year predictive model
Shot power
PDP jumping three-year predictive model
Jumping
PDP stamina three-year predictive model
Stamina
PDP strength three-year predictive model
Strength
PDP long-shots three-year predictive model
Long-shots
PDP aggression three-year predictive model
Aggression
PDP interceptions three-year predictive model
Interceptions


PDP positioning three-year predictive model
Avg. pred. performance
Positioning
PDP vision three-year predictive model
Vision
PDP penalties three-year predictive model
Penalties
PDP marking three-year predictive model
Marking
PDP standing tackle three-year predictive model
Standing tackle
PDP sliding tackle three-year predictive model
Sliding tackle
PDP goalkeeping diving three-year predictive model
Goalkeeping diving
PDP goalkeeping handling three-year predictive model
Goalkeeping handling
PDP goalkeeping kicking three-year predictive model
Goalkeeping kicking
PDP goalkeeping positioning three-year predictive model
Goalkeeping positioning
PDP goalkeeping reflexes three-year predictive model
Goalkeeping reflexes
PDP attacking work rate three-year predictive model
Low
Medium
High
Attacking work rate

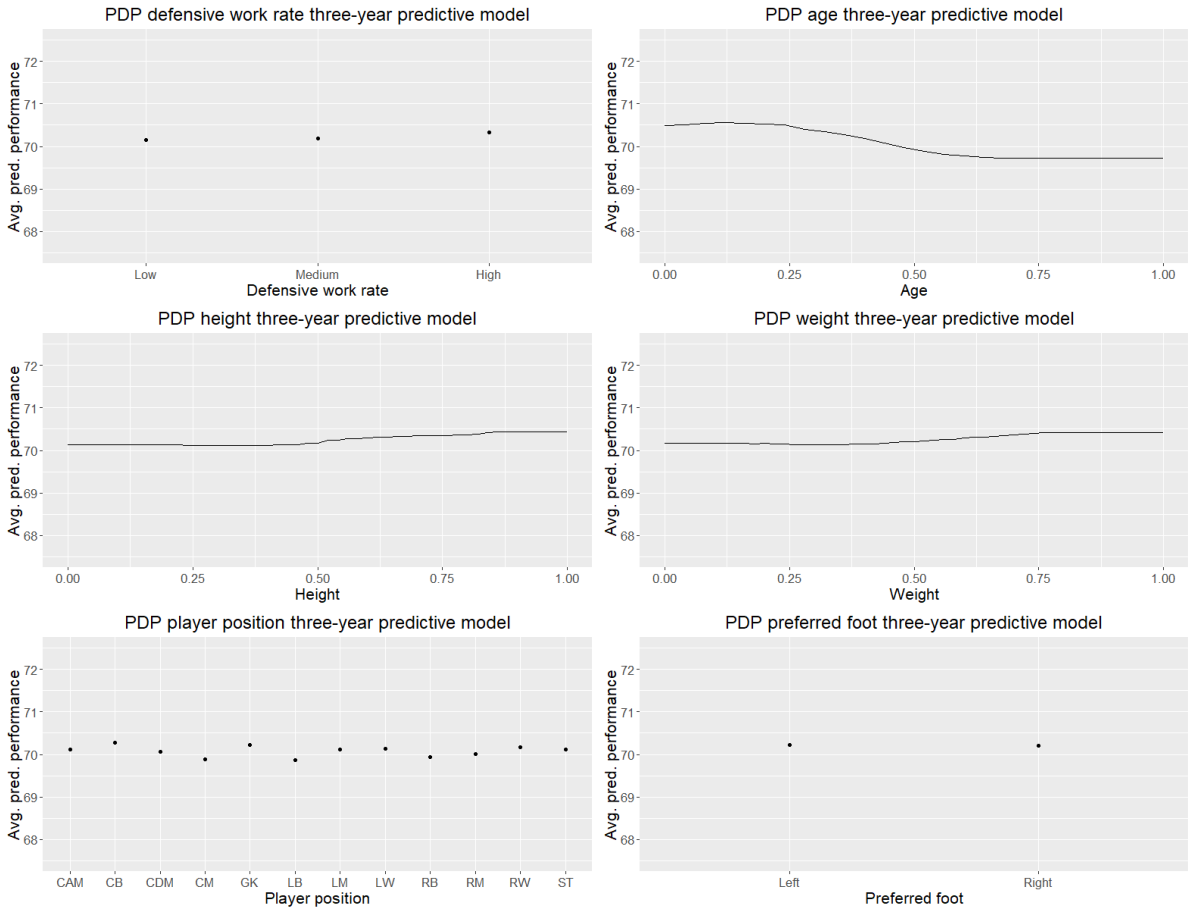
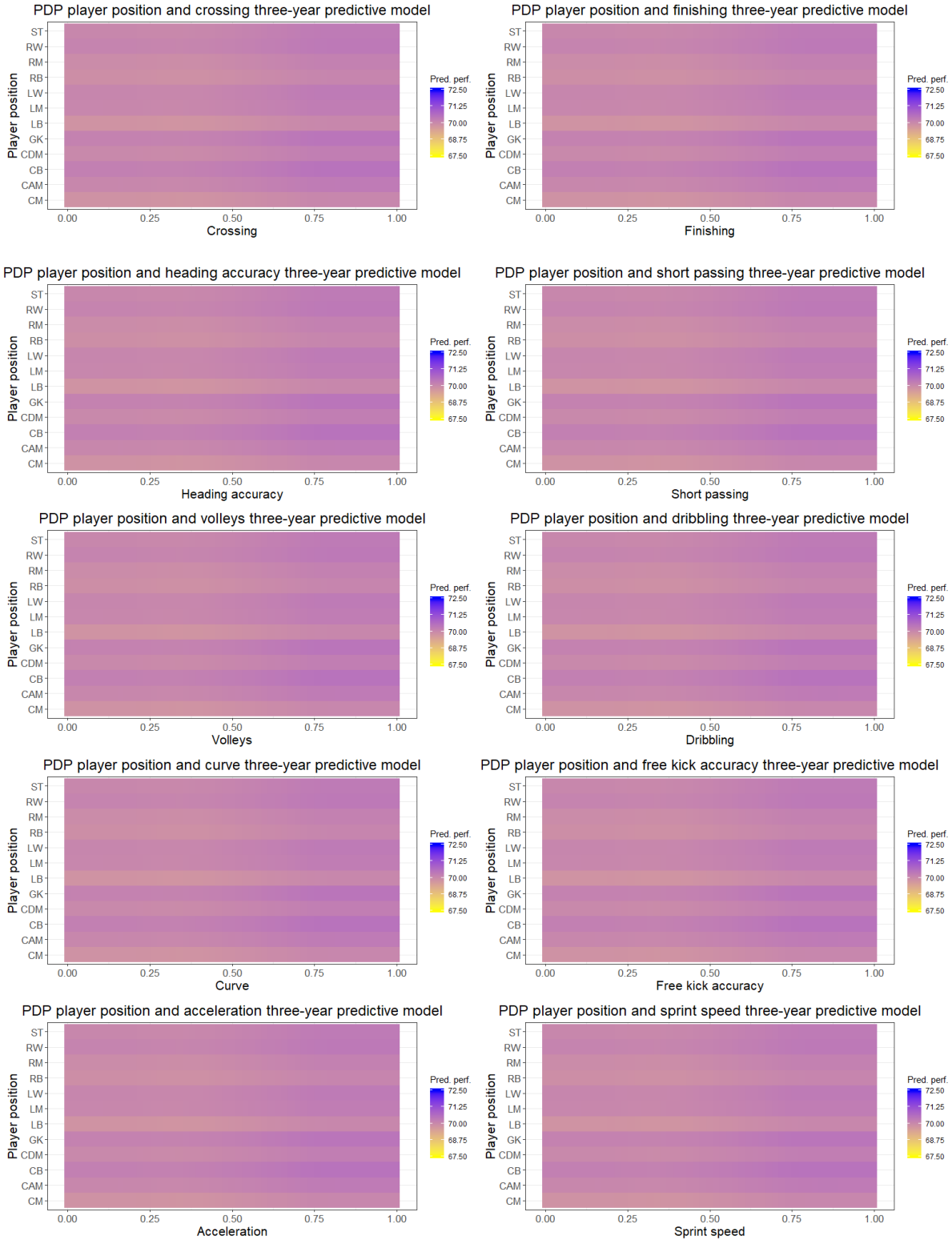
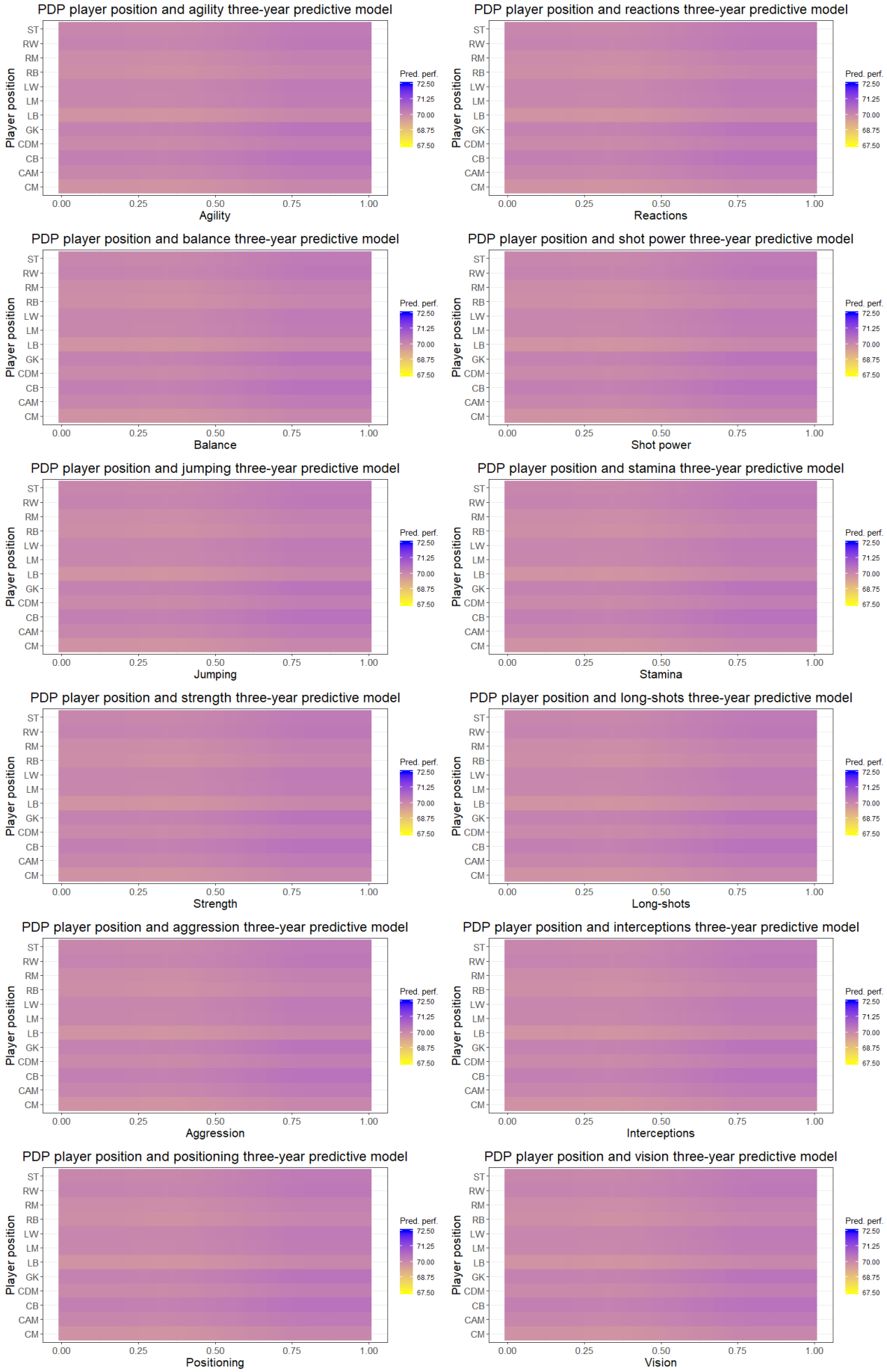
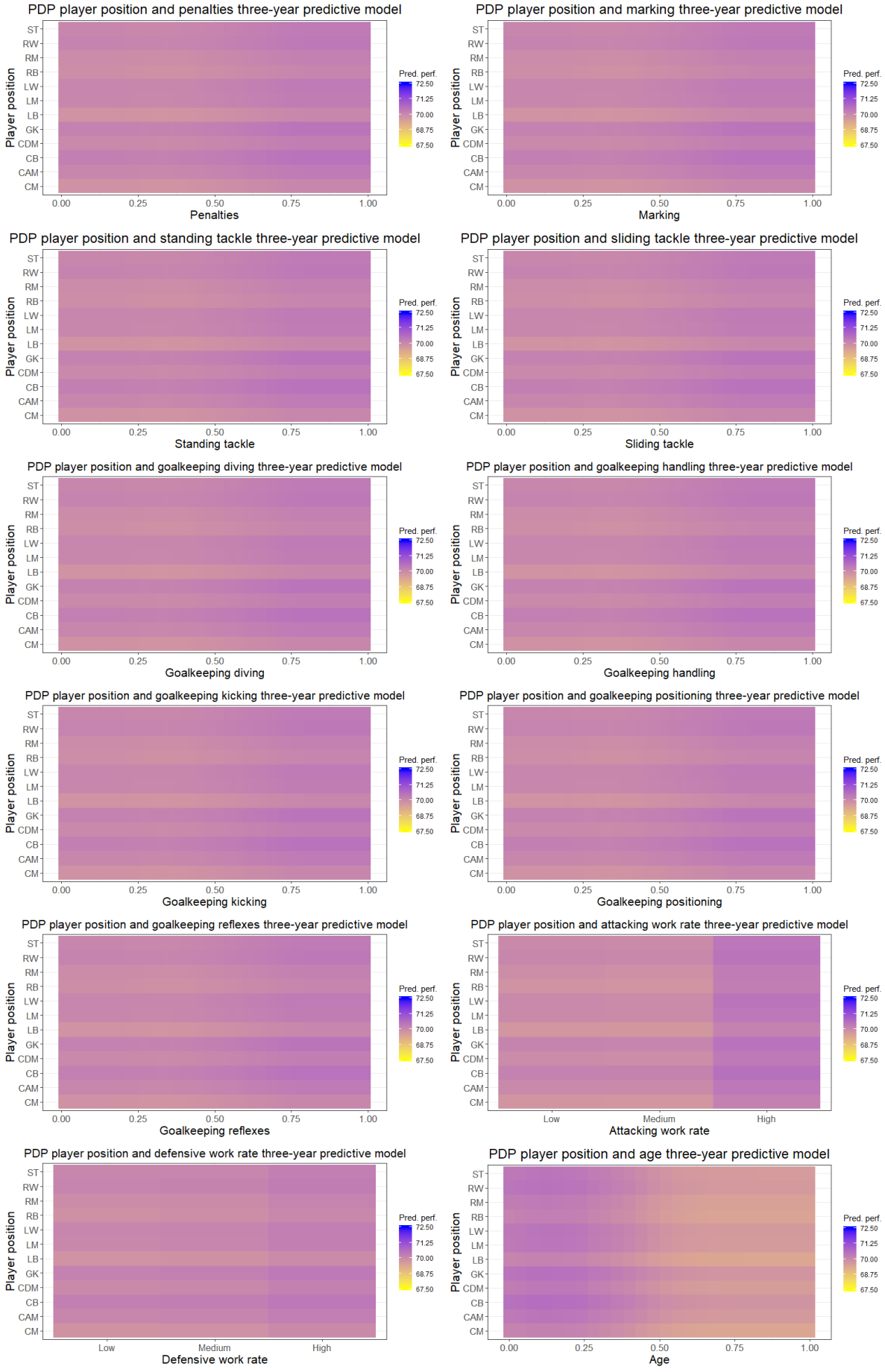


Figure 23: Partial Dependence Plots of all variables in the three-year predictive Random Forest

16. Appendix H: Partial Dependence Plots of interaction effect of player position in three-year predictive Random Forest




PDP player position and agility three-year predictive model
PDP player position and reactions three-year predictive model
PDP player position and balance three-year predictive model
PDP player position and shot power three-year predictive model
PDP player position and jumping three-year predictive model
PDP player position and stamina three-year predictive model
PDP player position and strength three-year predictive model
PDP player position and long-shots three-year predictive model
PDP player position and aggression three-year predictive model
PDP player position and interceptions three-year predictive model
PDP player position and positioning three-year predictive model
PDP player position and vision three-year predictive model
Player position
ST
RW
RM
RB
LW
LM
LB
GK
CDM
CB
CAM
CM
0.00
0.25
0.50
0.75
1.00
Agility
Reactions
Balance
Shot power
Jumping
Stamina
Strength
Long-shots
Aggression
Interceptions
Positioning
Vision
Pred. perf.
72.50
71.25
70.00
68.75
67.50


PDP player position and penalties three-year predictive model
PDP player position and marking three-year predictive model
PDP player position and standing tackle three-year predictive model
PDP player position and sliding tackle three-year predictive model
PDP player position and goalkeeping diving three-year predictive model
PDP player position and goalkeeping handling three-year predictive model
PDP player position and goalkeeping kicking three-year predictive model
PDP player position and goalkeeping positioning three-year predictive model
PDP player position and goalkeeping reflexes three-year predictive model
PDP player position and attacking work rate three-year predictive model
PDP player position and defensive work rate three-year predictive model
PDP player position and age three-year predictive model
Player position
ST
RW
RM
RB
LW
LM
LB
GK
CDM
CB
CAM
CM
Pred. perf.
72.50
71.25
70.00
68.75
67.50
0.00
0.25
0.50
0.75
1.00
Penalties
Marking
Standing tackle
Sliding tackle
Goalkeeping diving
Goalkeeping handling
Goalkeeping kicking
Goalkeeping positioning
Goalkeeping reflexes
Low
Medium
High
Attacking work rate
Defensive work rate
Age

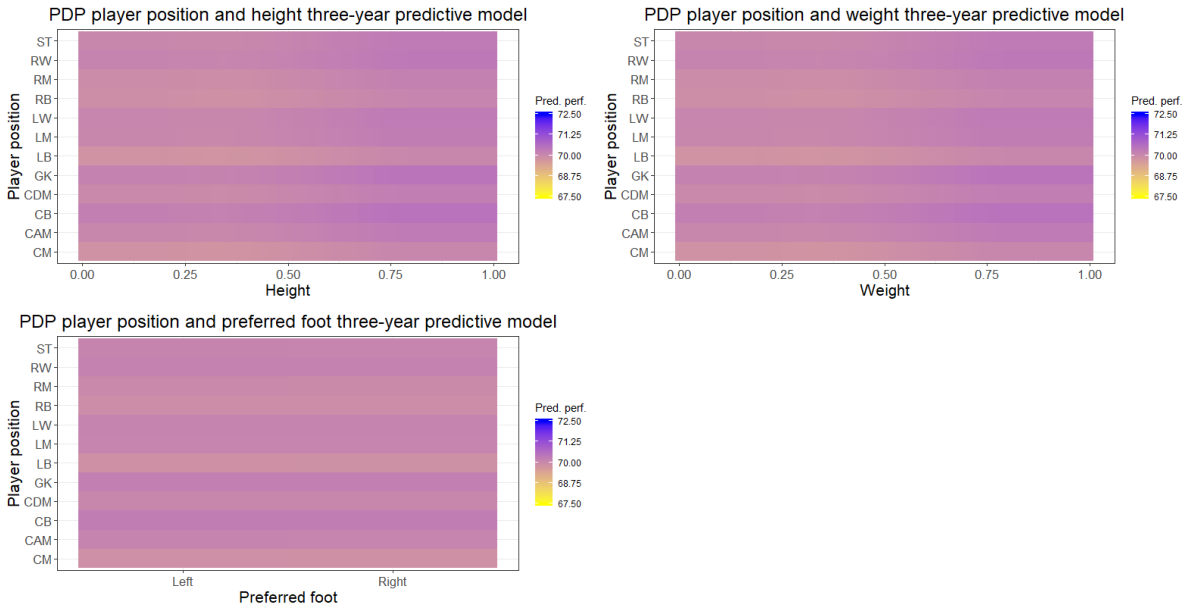
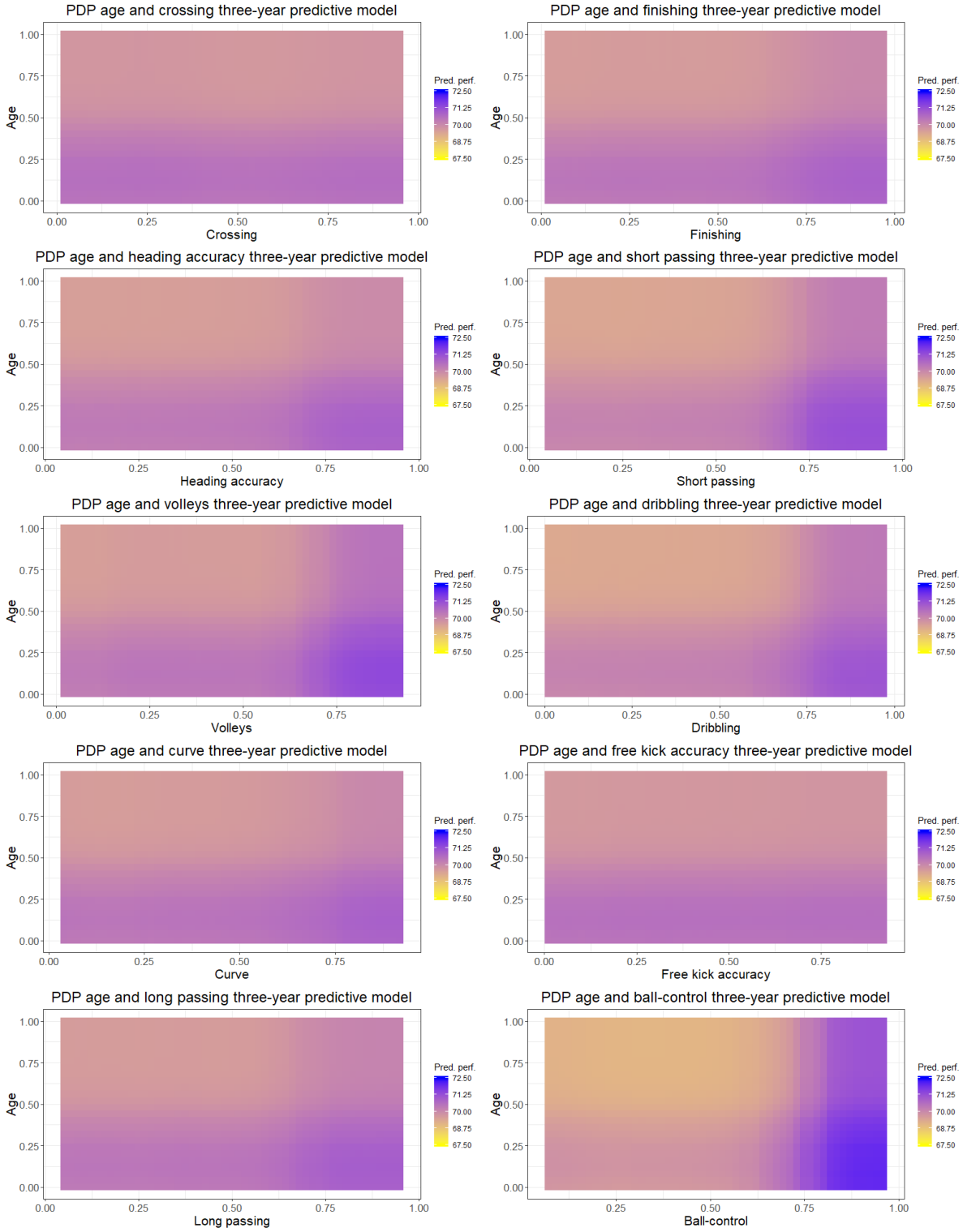
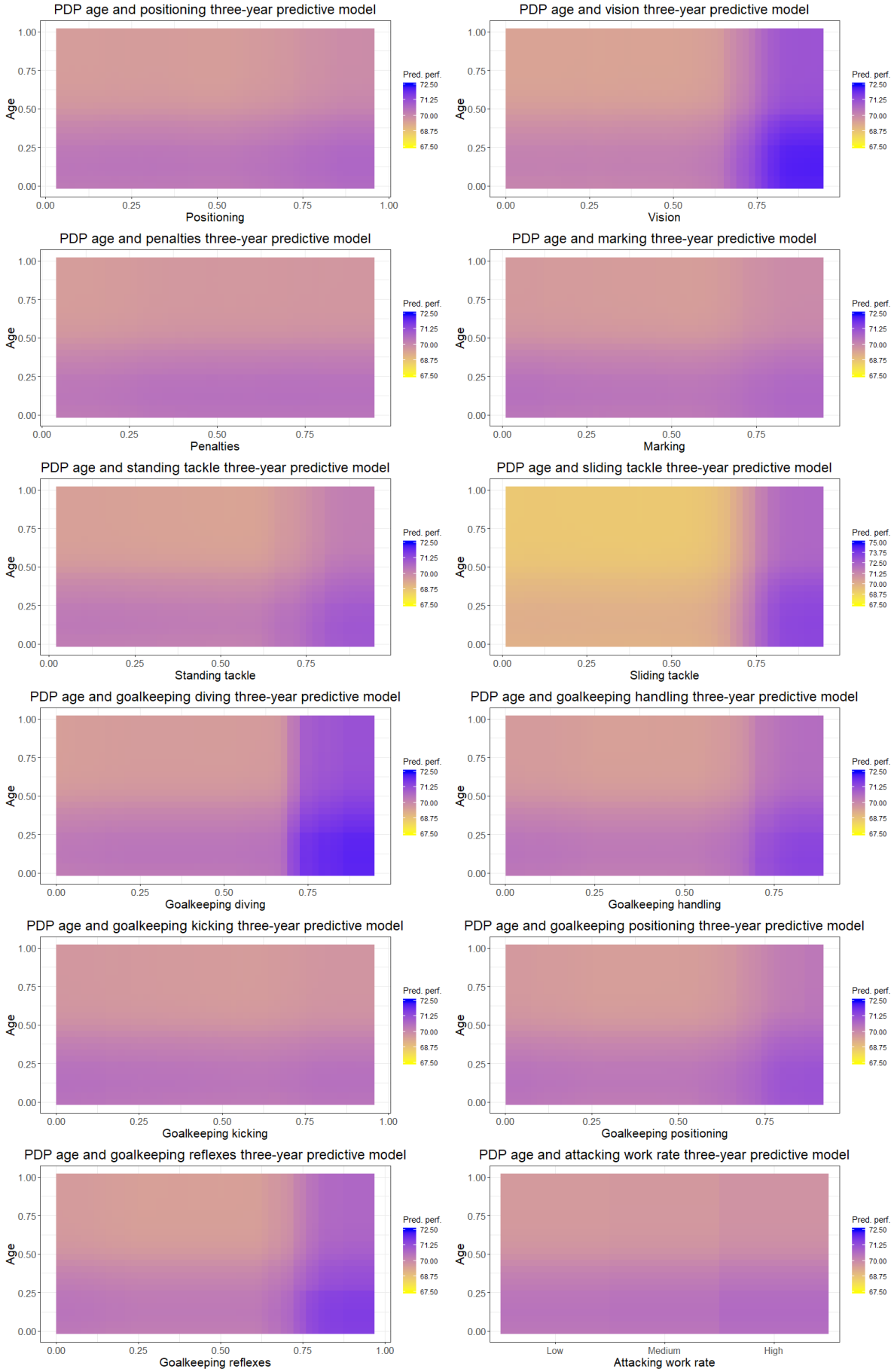


Figure 24: PDPs of all variables in the three-year predictive Random Forest with player position included as interaction effect

17. Appendix I: Partial Dependence Plots of interaction effect of age in three-year predictive Random Forest




PDP age and positioning three-year predictive model
Positioning
PDP age and vision three-year predictive model
Vision
PDP age and penalties three-year predictive model
Penalties
PDP age and marking three-year predictive model
Marking
PDP age and standing tackle three-year predictive model
Standing tackle
PDP age and sliding tackle three-year predictive model
Sliding tackle
PDP age and goalkeeping diving three-year predictive model
Goalkeeping diving
PDP age and goalkeeping handling three-year predictive model
Goalkeeping handling
PDP age and goalkeeping kicking three-year predictive model
Goalkeeping kicking
PDP age and goalkeeping positioning three-year predictive model
Goalkeeping positioning
PDP age and goalkeeping reflexes three-year predictive model
Goalkeeping reflexes
PDP age and attacking work rate three-year predictive model
Attacking work rate
Age
Pred. perf.
Low
Medium
High

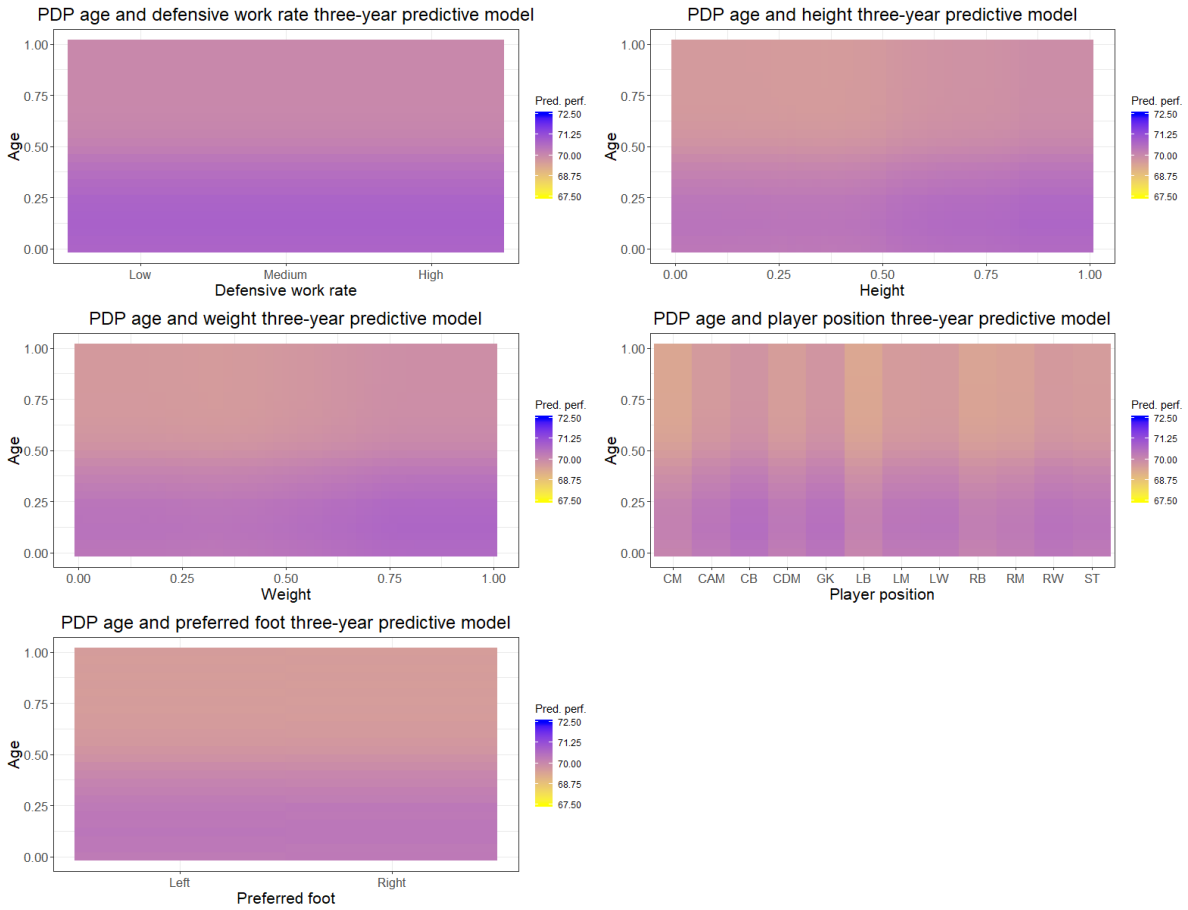


Figure 25: PDPs of all variables in the three-year predictive Random Forest with age included as interaction effect