Vrije Universiteit Amsterdam

Master Thesis

Predicting performance using player characteristics

Author: Jack Spanjer (2643915)

1st supervisor:	Rob van der Mei	
daily supervisor:	Noosheen Gholami	Forward Football
2nd reader:	Francois Lavet	

A thesis submitted in fulfillment of the requirements for the VU Master of Science degree in Business Analytics

September 4, 2023

Preface

This thesis is written as graduation project for the Business Analytics Masters program at the Vrije Universiteit (VU) located in Amsterdam. The thesis is a report of a six month internship at Forward Football, a company that helps football clubs/organizations improving their football performance using high-end data analysis. The aim of this thesis is to help them expand their knowledge in the improvement of talent identification/development.

I chose to apply at this company because of my passion for football, the opportunity to be able to work with professional football clubs and conduct tests on their youth players was truly an enjoyable experience. So I would like to thank the team at Forward Football for giving me the opportunity to write my thesis here, and in particular my company supervisor Noosheen Gholami for her input and advice. Finally, I would like to thank Rob van der Mei for being my supervisor from the VU and Francois Lavet for his willingness to be the second reader.

Management summary

Context. In the heavily competitive world of professional football, clubs strive to gain a competitive edge adopting new techniques to get an advantage on the field and consequently reaping the financial benefits. Especially for clubs without substantial financial resources, identifying talent in young players and developing them is a key component in building a sustainable foundation to maintain high-level performances.

Goal. The goal of this study is to capture important characteristics and skills of young players to be able to perform well in matches, gaining new insights in the field of talent identification. Leading to the research question: 'To what extent can player characteristics and/or skills be used to predict the performance of football players?'

Method. In the process of predicting a player's performance, fair metrics have to be used that not only use success rates, but also capture difficulty and effectiveness of the passes/dribbles/interceptions. Classification models are used to predict the difficulty of the passes and dribbles (xSP and xSD) using contextual features extracted from the coordinate data, which are used to calculate the passing and dribbling performance of the player. While for interceptions the frequency, defensive gain and offensive gain are used.

Results. The results show that the xSP could be very accurately estimated using a random forest. The best calibration of xSD models obtained using catboost was slightly worse, but still accurate enough to fairly assess a player's dribbling skills. The most important features in these models aligned with the expectations, the position on the pitch and pressure mainly influenced the xSD and xSP. In addition, the length and angle of the pass are also crucial features for xSP estimation. Adding test data as features only introduced noise and decreased model performance. Using regression models, the entire disparity in passing/dribbling performance of the players could not be predicted, but correlations were observed providing information about important characteristics/skills for players to have.

Conclusions. For passing, mental/cognitive features were most important, while for dribbling, biological age and dribble speed were the best indicators. XGBoost and KNN were the best for passes and dribble performance respectively, while interception performance could not be predicted well.

Contents

Glossary

ADASYN Adaptive Synthetic Sampling 21

ANN Artificial Neural Network i, 29

APHV Age of Peak Height Velocity 11, 44

DPR Dribbling Performance Ratio 2, 33, 35

ECE Expected Calibration Error 33

EFB Exclusive Feature Bundling 27

FCM Four Corner Model 3, 8

FST Football Skills Track 12

GID Game Insight inDicator 13

GOSS Gradient-based One-Side Sampling 27

IJBT Inside Joy Ball control Test 13

KNN K-Nearest Neighbors i, 28

KS Kolmogorov-Smirnov 17

LSPT Loughborough Soccer Passing Test 12, 13, 58

MAE Mean Average Error 33

MCAR Missing Completely At Random 16

MLP Multi-Layer Perceptron 29, 30, 31, 32

MSE Mean Squared Error 31

PCA Principal Component Analysis 18

PPDA Passes Allowed Per Defensive Action 7

PPR Passing Performance Ratio 2, 15, 16, 17, 18, 33, 35

PUP10 Passing Under Pressure 10m+ 13, 42

RMSE Root Mean Squared Error 33

ROS Random OverSampling 21

SMOTE Synthetic Minority Oversampling Technique 21, 25

TVPS Test of Visual Perceptual Skills 13

VFMT Visual Fine Motor Test 14
VIF Variance Inflation Factor 18

xG Expected Goals 5, 6
XGBoost eXtreme Gradient Boosting 5, 6, 27
xSD Expected Successful Dribbles ii, 21, 38, 39
xSP Expected Successful Passes ii, 2, 5, 21, 32, 35
xT Expected Threat 39

1

Introduction

1.1 Business context

Football is the world's most popular sport with over 4 billion worldwide viewers, consequently it is also the wealthiest in terms of generated revenue and expenditures. Due to the emotional connection fans have with the clubs and media coverage that mainly focuses on sporting performances, it is often forgotten that professional football clubs are businesses that engage in generating high revenue and manage costs while operating in a competitive market. To maximize sales of merchandise, tickets, broadcasting rights and sponsorship deals, clubs have to differentiate themselves from the competition using marketing and performance. Several studies have proven the positive correlation between the sports performances and financial health of the club (Alaminos *et al.*, 2020 (1), Miragaia *et al.*, 2019 (2)), meaning that success on the pitch is significant on both an emotional and financial level.

In professional football, matches are often decided in the smallest of details and decisions, even the slightest advantage can decide the outcome. This stresses the importance of careful preparation, smart investing and improve the technical, physical and mental skills of players to achieve success. According to Ratten, 2010 (3), innovation is a key aspect in having sustained success in the sporting world. Many clubs have already established dedicated data analytics departments to help gain competitive advantages. An integral part of achieving sporting and financial success lies in the quality of the youth academy for example, clubs such as Ajax and Benfica have both capitalized of their exceptional academies. Since 2015, both clubs sold over 250 million euros worth of youth players that led them to success in the Champions League further generating substantial financial benefits. To maintain high level performances and achieve financial stability, identifying and developing talent is crucial, in particular for clubs without substantial financial resources.

1.2 Research question

To gain new insights in talent identification and development, a deeper understanding of how specific player skills and characteristics translate to in-game performance has to be acquired. To achieve this, it will be researched whether the test data has predictive power of the in-game performances of youth players. The identification of key characteristics/skills that signify the best performing youth players will provide meaningful insights on talent identification and development. The central research question of this study is: 'To what extent can player characteristics and/or skills be used to predict the performance of football players?'

1.3 Research approach

To answer the research question several machine learning models will be trained and evaluated on general player information, test data and match data obtained at academies of over 10 Dutch and Belgian football clubs. Before predicting performances, it is essential that a suitable way of evaluating a player's performance is established using fair metrics. Completion percentage is a commonly used metric but is considered simplistic and ignores important factors such as context and difficulty. While the ability to perform difficult events is an indicator of raw talent regardless of the effectiveness. This will be done using classification models that estimate the difficulty of events. The probability that a pass is successful, the Expected Successful Passes (Expected Successful Passes (xSP)), will be predicted for each pass for pass difficulty estimation. Initially, a baseline is set using only contextual features of the pass. Subsequently, the test results are added as features for xSP prediction, aiming to improve the performance of the models. As the true success probability of the events is not known, the Brier score, Spiegelhalters Z-statistic, expected calibration error and calibration plot are used for evaluation. Analyzing how the test results influence predictions gives information on important characteristics of players, providing insights into crucial skills for specific pass types in all areas of the field. Next, regression models are applied to predict the overall passing performance of the players using the Passing Performance Ratio (PPR), which is the ratio between the number of completed passes divided by the expected completed passes of the player. The PPR is calculated using the xSP values estimated by the baseline classification model, otherwise players with good test performance will be punished in the overall performance assessment. The same approach will be applied for evaluating the dribbling skills using the Dribbling Performance Ratio (DPR). As interceptions are always successful, a different way of assessment has to be used. The metric used is a weighted average of a few selected features; frequency and the defensive/offensive value of the interception.

The conducted tests are based on aspects of The Football Association's Four Corner Model (FCM), developed by the FA for assessment of youth development.

1.4 Organization

This paper consists of five main chapters, Chapter 2 analyzes similar researches in the field of study to provide the current state of knowledge and establish what this paper adds to this. Chapter 3 provides insight in the data that is used for this study and how this data is prepared for machine learning. Chapter 4 describes the methods used to obtain the results; the respective results will be discussed in Chapter 5.

2

Literature Review

Due to the still growing popularity and increasing financial interests, the use of data science in football has seen a marked increase over the last decade. This is observed in practice, as an extensive amount of research has been conducted over the years and currently data science is a prominent aspect of the sport.

2.1 Research on expected successful passes

Researching which characteristics influence performance using test data can be approached using **classification** and **regression models**. Classification will be used to see whether test data improves the xSP model. If certain abilities are important to complete specific passes, adding test data can lead to an improvement in the results of the xSP model. Passing is one of the most important aspects of football. However, little research has been conducted on the topic. It is often observed that players are compared or judged on their passing completion percentage, ignoring difficulty and quality of passes. In a recent interview with Sky Sports, Manchester City's star midfielder Kevin de Bruyne expressed his opinion on the metric stating, 'Pass completion is one of the most wasteful stats, it doesn't define me as a player. You can have 96% pass completion, but if I play it sideways or backwards I don't create anything.' (4). Few papers did research on the prediction of the likelihood of pass completion, Anzer and Bauer (5) used eXtreme Gradient Boosting (XGBoost) with 25 contextual features of the pass and divides the predictions into three levels of pass difficulty. As no true label is known, experts manually label the level of difficulty of the passes. The models are then evaluated based on expert accordance, achieving 78%. Unfortunately, manual expert labeling is not available in this research. With this model they argue that a player's passing performance can be assessed more accurately using risk profiles and efficiency.
Significantly more research has been conducted on Expected Goals (xG), which is very similar to xSP meaning that studies researching xG contain valuable information for this study. For example, Fairchild *et al.* (6) predicted probabilities for a shot to be a goal

based on contextual features of the shot, which is what this paper strives to achieve for passes. The paper by Fairchild *et al.* models a sequence of shots through a Poisson binomial distribution, as all shots are taken independently and not necessarily follow the same distribution. Using logistic regression, they predict whether a shot results in a goal or not. However, instead of the usual classification metrics such as accuracy, recall and precision, they propose a different way of evaluation. Considering the number of correct predictions does not accurately reflect whether the probabilities of the shots are accurate. They argue that two models that predict 60% and 85% for a shot respectively, have the same outcome as both predict that the shot is a goal, while it is not known which model is more accurate. The accuracy of the predicted probability should be the main evaluation criterion, not the accuracy in predicting whether a shot results in a goal. They propose a calibration curve, for a set of shots with probabilities in a certain range around probability π, then π of those shots should result in a goal. A model with accurate predictions would have the line x=y within the linear fit and its confidence interval. The *Brier score* β is proposed as second evaluation metric, which can be calculated using Equation 2.1. With N being the number of observations, π_i is the predicted probability of instance i, and y_i is the label of instance i. Fairchild *et al.* achieved that the x=y line is within the 95% confidence interval of the linear fit and a Brier score of $\beta = 0.19$. Where β is defined as:

$$\beta = \frac{1}{N} \sum_{i=1}^{N} (\pi_i - y_i)^2 \tag{2.1}$$

A similar study conducted by Pardo (7) also creates models that predict the xG and similar to this paper, it investigates whether adding player quality to the contextual information improves the models. However, it does not use data collected by conducting tests, but the player statistics of the FIFA video game by EA Sports. Using logistic regression, a random forest and a neural network, it is concluded that adding player quality yields a slight improvement. After the calibration plot like Fairchild, several post-processing calibration methods such as No post-calibration, Isotonic regression and Platt scaling, are used to further improve the results. As the original calibration was already accurate, the post-processing calibration methods had no significant effect.

Similar to Anzer and Bauer, Cavus and Biecek (8) use Gradient Boosting models to predict xG. Instead of only using XGBoost, they also implement different algorithms from the gradient boosting family in LightGBM and CatBoost. The three gradient boosting algorithms had very similar performances which were slightly worse than the random forest. Several papers have compared the three variations of the gradient boosting algorithm, however there is no algorithm that always outperforms the others (9)(10)(11).

2.2 Evaluation of dribbling

Like evaluating the passing abilities of a player, the success rate is not an accurate metric for a player's dribbling skills. A paper by Dick *et al.* (12) evaluates the player's situative dribble effectiveness by looking at the increase in 'V', which is a number that illustrates the likelihood of success of the team. This value is indicates how the player increases the chance of scoring based on his dribble. This metric favors attacking midfielders and fullbacks considering they are often the ones to bring the ball up the field. They rate the players based on the number of dribbles per minute that have a higher gain than a static threshold of V, setting the importance of medium dribbles on the same level as higher quality dribbles. Decroos *et al.* (13) uses the difference in scoring probability to value a players action, given by Equation 2.2. Where the value of action a at time i is the difference in scoring probability in game state S_i ($P_{\text{scores}}\ (S_i, x)$) and the probability to score in the previous game state S_{i-1} ($P_{\text{scores}}\ (S_{i-1}, x)$). Their dataset is very similar to the one used in this study, featuring passes (64.63%), dribbles (8.69%), and interceptions (5.01%) as event types.

$$\Delta P_{\text{scores}}\ (a_i, x) = P_{\text{scores}}\ (S_i, x) - P_{\text{scores}}\ (S_{i-1}, x) \tag{2.2}$$

Considering these papers base dribble quality purely on the increase in chance to score, it is difficult for attackers to highly contribute while performing dribbles under the highest pressure. Only using this feature ignores dribble context. It could be argued that players with a higher attacking contribution are better performers than players who are able to perform harder dribbles. However, especially for young players, the ability to complete difficult dribbles (or passes) is an indicator of raw talent. With the right coaching and development, decision making can be improved and this skill can be refined to be used more effectively. Hence, this paper will develop a different metric that estimates the difficulty of dribbles and passes.

2.3 Evaluation of interceptions

The evaluation of interceptions is slightly different compared to passing and dribbling. As an interception is always a successful ball recovery, the frequency becomes more important as it indicates the ability of the player to consistently recover the ball. The frequency is often expressed as interception per x time units (often 90 minutes). However, Trainor *et al.* (14) introduced a metric to measure a players involvement in defensive actions based on the number of potential interceptions called Passes Allowed Per Defensive Action (PPDA), calculated using Equation 2.3. This is mainly used to evaluate a team's pressing style, but can be adjusted to get a PPDA value for each player.

$$\text{PPDA}_A = \frac{\text{Number of passes }_B}{\text{Number of defending actions }_A} \tag{2.3}$$

Using this metric for player evaluation has several drawbacks, players in defensive minded teams will be heavily penalized for the playing style of their respective teams. Only focusing on frequency also ignores the importance of interceptions. An interception when you're the last defender is significantly more important than an interception on the midfield. Merhej *et al.* (15) researched using deep learning by estimating how much threat is prevented.

A masters thesis by Piersma (16) further improved defensive action evaluation by incorporating interception difficulty and missed opportunities. This is done by examining all game states using machine learning on a data set with over ten million events. Unfortunately, due to time and data constraints it is not possible to create a similar model for defensive action valuation.

2.4 Translating test data to in-game performance

Predicting the overall performance of a player can be seen as a regression problem, it is seen that several types of regression models were used in similar studies. An example of this is a paper written by McGuckian et al. (17) which uses subset, linear and non-linear regression models to identify the association between visual exploration and passing performance. They concluded that a high number of head turns before the ball was received heavily correlated with good passing performances. Other studies that also perform tests on players and connect them to in-game performances often only examine single correlations of a test and a game performance indicator instead of combining test results to make predictions. Examples of this are papers by Bila and Hillman (18) and Lipinska and Szwarc (19). Lipinska and Szwarc used Spearman's, gamma and Kendall's tau rank correlation coefficients to prove a correlation between physical components and match performance. They found that good performance on speed tests and leg power/strength were significantly correlated with one-on-one dribbling success. Bila and Hillman show a relationship between the mental skills of a player and their short-passing performance. Separate Spearman's rho values were used to identify the relationship between tests and anxiety/confidence. However, Abdullah et al. (20) have shown that even though there is a correlation between psychological factors and performance, psychological factors alone could not accurately predict the performance of a player. Thus, even though correlations can exist between a test result and part of a player's performance, one part of a player's skill set is not enough to accurately predict the overall performance of a player.

A study which also assesses youth players through the four categories of the Four Corner Model is written by Kelly *et al.* (21). Cross-validated Lasso regression on test and match data was used to predict subjective performance, which are grades given by players and coaches, and the likelihood of signing a pro-contract. Fifteen features showed a non-zero

SD.

2.5 Talent identification

The field of talent identification and development has been widely researched over the years, Sarmento *et al.* (22) have done a systematic review of the performed studies in the field. They concluded that coaches should consider scaling a player's technical, tactical and physiological skills against age. Furthermore, the current knowledge lacks information on the influence of psychological and environmental factors. According to Fortin-Guichard *et al.* (23), the most common approach to identify talent indicators is to compare the differences between selected and de-selected players while the selection itself still remains a subjective process decided by coaches. They also state that current studies often ignore the biological age of the players, as some humans mature earlier than others this can lead to overlooked potential in players that mature at a later age.

This study tries to bridge these gaps by capturing the differences in test results by comparing players based on their objectively measured match performances. Regression models are applied to predict passing, dribbling and intercepting ratings that are calculated with probabilistic models. The regression models use features based on twelve tests that, contrary to most other papers, examine the entire skill set of a player by measuring technical, physical, psychological and cognitive skills, combined with a player's biological age.

3
Data

The data has been acquired by Forward Football through conducting tests and collecting match data at twelve clubs in the Netherlands and Belgium. Throughout the six months, visits have been made to these clubs, actively participating in the data collection process. In this section, all information regarding the data will be provided. First, in subsection 3.1, a general explanation of the data used in this research will be provided. Then, the data will be prepared and explored in section 3.2. Information about the meaning of all features and their respective ranges is seen in Tables A.1, A.2 and A.3 in the Appendix.

3.1 Data description

In this section, general information about the available data is given. Based on sub-scores of the tests, characteristics of players will be computed according to Forward Footballs guidelines. For a more detailed description of all features, Tables A.1, A.2 and A.3 in the Appendix give an oversight of all constructed features, by providing information about the definition and the range of values observed in the data set.

3.1.1 Player information

The first part of the database contains the player profile and physical characteristics. The player profile contains the professional information of the players such as their position, club, team and KNVB ID. The physical characteristics mainly include height, weight, age and their Age of Peak Height Velocity (APHV), which is the age a person reaches their maximum growth rate.

3.1.2 Test performance

Twelve tests have been conducted on players from youth teams of twelve football clubs, which can be divided into five categories. These tests provide information on the technical, cognitive and physical abilities of a player. Physical abilities are integrated in the tests of the other categories. The fourth and fifth categories are questionnaires about the mental strength and sport history of the player. As the procedures of the tests are not common knowledge, the conducted tests will be briefly explained to provide insight on what the data means and how it is utilized in the prediction of the performance of the players. Unfortunately, none of the players have participated in every test.

3.1.2.1 Technical tests

The first technical test conducted is the 'Football Skills Track' (FST), which assesses the player's proficiency in dribbling. The track consists of five sections with different dribbling styles, the cumulative time the player needs to complete these sections is their FST test result. The whole track is depicted in Figure 3.1.

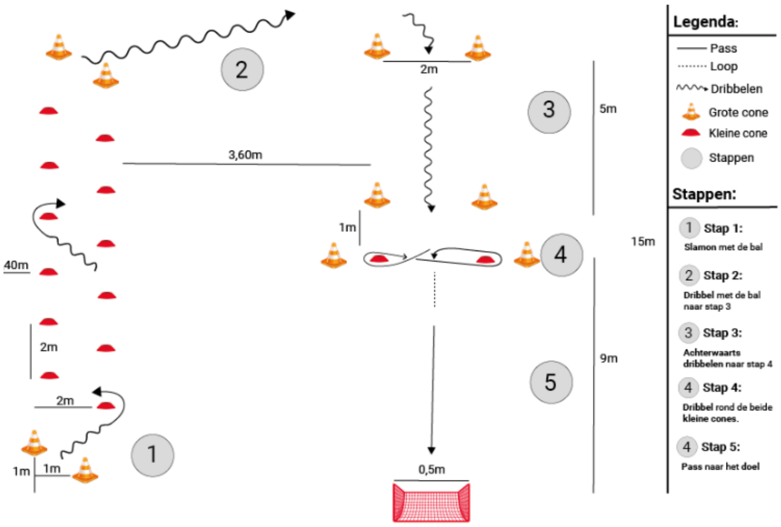

Figure 3.1: Schematic overview of the Football Skills Track (24)

The second technical ability evaluated is short passing, which is done with the Loughborough Soccer Passing Test (LSPT). BenOunis et al. (25) proved the effectiveness of the LSPT and that it can be used to distinguish elite players from less skilled players. The player passes sixteen times to benches as fast as possible while minimizing penalty points. The total time plus penalty points is the final score of the player. A schematic overview of the test is depicted in Figure 3.2.

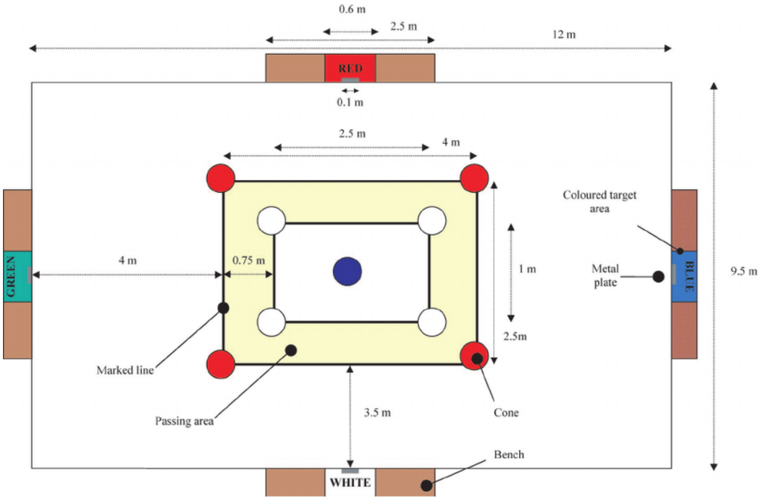

Figure 3.2: Schematic overview of the Loughborough Soccer Passing Test (26)

Then, the speed and fluency of the ball control are tested with the Inside Joy Ball control Test (IJBT). The test measures the fluency and quantity of six separate exercises. The final score is calculated as the normalized average of the quantities.

The last technical test will test the long passing under pressure ability. During the Passing Under Pressure 10m+ (PUP10) test, players have to pass over eleven and sixteen meters in two directions using both feet while a defender puts pressure on them. Penalty seconds will be given when a pass is inaccurate. The final score is the total time plus the penalty seconds.

3.1.2.2 Cognitive tests

Cognitive abilities such as decision-making, spatial awareness, reaction time, and focus are crucial for players to perform well on the field. Four tests are developed to assess several of these cognitive abilities. The first cognitive test is the Game Insight inDicator (GID) test, which tests insight, reaction time, positioning and judgment of the players. Six zones are presented to the player, divided in three columns and two rows. The objective of the test is to estimate in which zone the ball will end up based on video footage that turns black before, during, or after the final ball contact of a player. If the player chooses the right row, one point is awarded, a second point if the right zone is chosen and three if it is done within the designated time which differs per situation. Every player does this for ten different scenarios.

Then, spatial perception and visual memory are tested with the Test of Visual Perceptual Skills (TVPS). The TVPS score is an average of five different tests which measure the visual spatial relations, memory, sequential memory, closure and figure ground. During the exercises the individuals have to remember, recognize and spot differences in figures.

Next, the eye-foot coordination of the players is assessed using the Visual Fine Motor Test (VFMT). The players are required to complete a speed ladder with agility cones as quickly as possible, every time a cone is touched the player has made an error. The calculation of the final score is given by Equation 3.1

$$\text{TVPS score} = \text{endtime} * (20 : (20 - \text{number of errors})) \tag{3.1}$$

Finally, the focus and information processing speed is tested using the RightEye computer which measures the eye movements of the player during various games.

3.1.2.3 Questionnaires

The players have participated in two questionnaires, one evaluates the mental strength of the player, and the other collects information on the player's sports history. The mental strength is assessed based on statements about mindset, setting goals and mental toughness, where the answers are numbers between 1-5 that indicate the level of agreement with the statement.

The sports history questionnaire asks about what sports the player has played and the number of hours practiced during each age. The questionnaire is based on a research from Côté *et al.* (27) that categorizes practicing sports in four categories; free play, deliberate play, structured practice and deliberate practice. For optimal talent development, they propose the framework depicted in Table 3.1. At a young age, an athlete is in their try out phase where the emphasis is on trying a wide variety of sports without structured training. From age thirteen, the athlete has to start specializing in a sport and balance structured training with other activities. From age 16 the athlete has to be fully invested in the sport and full focus has to be placed on structured training.

Phase	Deliberate play + other sports activities	Deliberate practice	Number of sports
Try out (6-12)	80%	20%	3-4
Specialization (13-15)	50%	50%	2-3
Investment (16+)	20%	80%	1-2

Table 3.1: Talent development framework proposed by Cote *et al.* (27)

3.1.3 Match data

Data of several matches is available, every match has data of the coordinates of all players and the ball at a frequency of 5 times per second. Additionally, every event that takes place including passes, tackles/interceptions, shots and dribbles.

3.2 Data preprocessing and exploration

In this section the data will be prepared for machine learning together with data exploration to get an initial indication of the data. As these tasks are often intertwined, these will be done simultaneously.

3.2.1 Player profile data set

With the data described above, two distinct types of data sets will be made. First, the data set for regression is made, which contains a player profile including general information, all test results and the overall match performances for all players. As every test for each team is saved in separate Excel files, these have to be merged to one data set containing all test results for each player. To ensure that the files merge correctly, issues such as differences in formats, score calculations and column names for the same tests, name misspellings and other data inconsistencies have to be corrected. To further clean the data, columns only containing NaN values and rows with players that haven't participated in a match or a test will be dropped.

This results in a database of 630 players and 983 columns where most columns are sub-scores of tests or questions of the questionnaires. As this is not a desirable setup, the sub-scores will be combined to define a specific characteristic of a player. The (inversed) average answers of the mental questionnaire are also divided into several characteristics. Finally, for each player, the ratio of deliberate practice and other activities during each phase will be calculated using the sports history questionnaire. Reducing the total columns to just 60 without losing information.

The PPR, DPR and interception rating mentioned serve as the target values of the regression models. A more in-depth explanation including mathematical formulations of these performance metrics can be found in Subsection 4.2.3.

As initial data exploration, Figure 3.3 depicts the individual correlations of the numerical predictor variables and the PPR. It is observed that no excessively strong linear correlations are observed, the sum of the absolute correlations is 2.1, further confirming the absence of strong correlations. Even though no initial strong correlations are found, it could still be the case that non-linear correlations exist that are not immediately observed.

3.2.1.1 Data imputation

As no player has participated in every test, most columns consist of more than 30 percent of missing values. To get a better understanding of the sparsity of the data, Figure 3.4 depicts for each column the percentage of data that is missing.

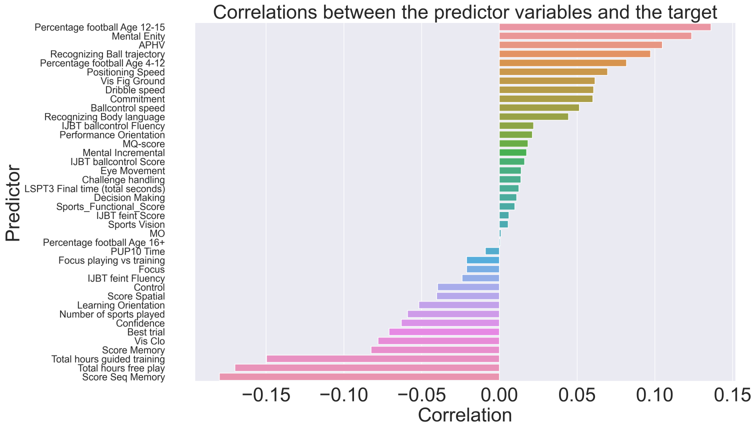

Figure 3.3: Correlation of predictors and PPR

Considering that the reason some players did not perform specific tests is independent of the characteristics of the player, the missing data is regarded as Missing Completely At Random (MCAR). Columns with over 80 percent missing values are removed from the dataset completely.

The numerical missing values will be imputed using the KNN imputer, as it is non-parametric meaning that it does not assume a specific distribution of the features. It is also able to handle more complex, non-linear relationships. The main drawback is that it is computationally expensive for large data sets, which does not apply here, since the data sets are not excessively large.

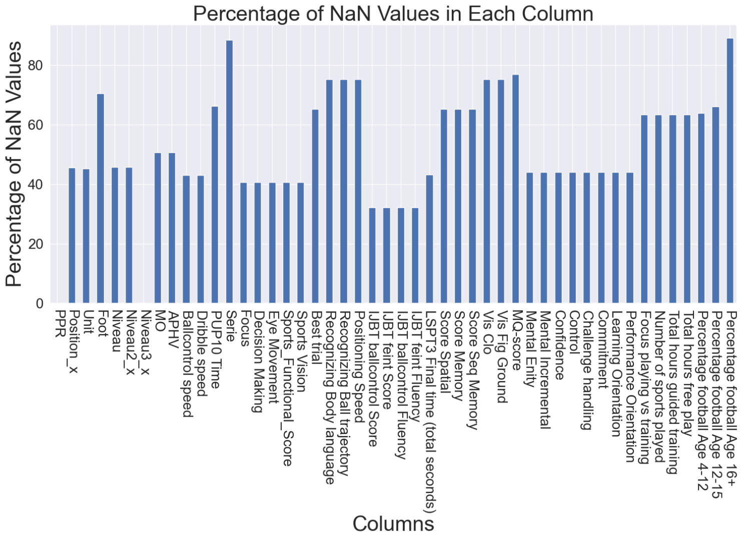

Figure 3.4: Percentage of missing data per column

Furthermore, according to Peng and Lei (28), KNN seems to perform better than CD, Mean/Mode Imputation and Multiple Imputation by Chained Equations because it is most robust to bias when the percentage of missing data increases. As the data set is sparse, this is an important feature. When using the KNN imputer, the optimal value for the k parameter has to be determined, which has been subject to controversy (29). Lall and Sharma (30) argue that $\sqrt{N}$ is the optimal value for k, with N being the sample size of the data set. Other studies (31) get a significantly lower optimal k of 10 with a data set larger than 100 samples. Due to the uncertainty in this topic, multiple values for k will be tested to see what yields better results. Due to the sparsity and the fact that for some tests there is no player who has done both, imputation was difficult. It is found that a k of $\sqrt{N}$ preserved the original correlations slightly better than the lower values for k, since it has an absolute correlation sum of 1.75 compared to 1.61. Figure 3.5 shows the results of imputed PUP10 values with k=5 against the PPR and Figure 3.6 shows the result for k = 25. Imputation with k = 5 causes many of the imputed values to have the similar values to outliers, which is illustrated by the vertical lines in the plot. This is less observed when using k = 25, which imputes closer to the mean while still keeping some variation.

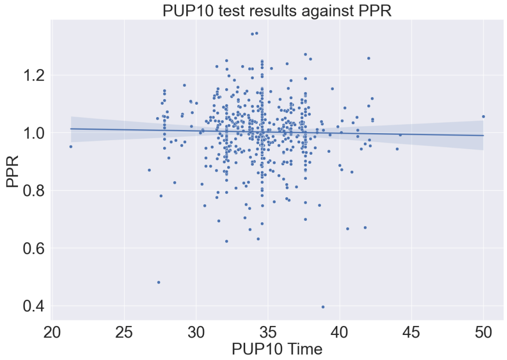

Figure 3.5: Imputation of PUP10 using k=5

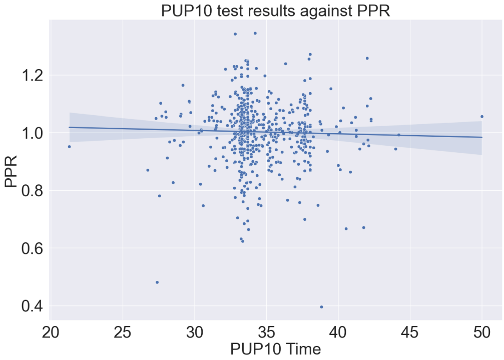

Figure 3.6: Imputation of PUP10 using k=25

3.2.1.2 Data transformation

Before performing further data pre-processing steps, it is important consider that the players face a different quality of opponents in the matches, matches are between teams consisting of players of the same age and level. It will be investigated whether the test results of players of different ages and different levels differ. The Kolmogorov–Smirnov (KS) test will be applied to compare all test distributions, as the KS test is efficient and non-parametric (30). The null hypothesis H_0 : 'The test results of O13 and O16 players have the same distribution', will be rejected at an $\alpha < 0.05$. 43 out of the 50 tests have rejected H_0, meaning that the test results follow a different distribution most of the time. It is observed that the tests which are not rejected are predominantly, but not exclusively cognitive tests. To illustrate the difference in distributions, Figure 3.7 shows the ball control speeds of O13 and O16 and it is clearly observed that O16 players are faster with the ball. While Figure 3.8 shows that the distribution of the eye movement is similar.

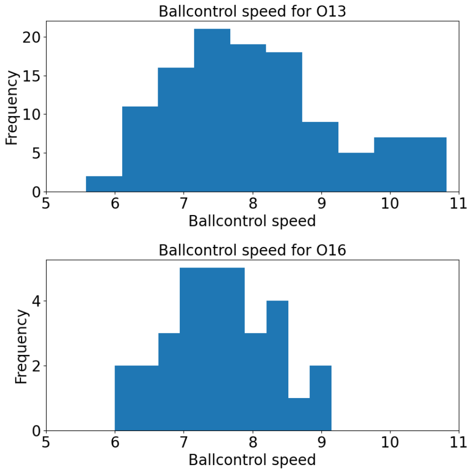

Figure 3.7: Ballcontrol speeds for O13 and O16

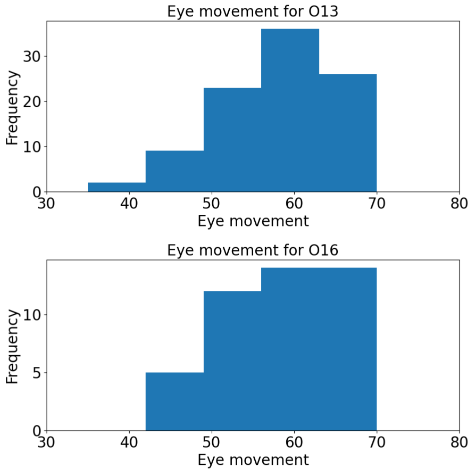

Figure 3.8: Eye movements for O13 and O16

This process is repeated for Sub-elite and Elite level players, where 47 of the 50 distributions are found to be different. However, the null hypothesis of the PPR feature cannot be rejected, which can be explained by the difference in opponent difficulty. To ensure that opponent difficulty is taken into account, the test data are scaled by age/level using the *StandardScaler* library of *sklearn*.

Before regression can be done, it is important to check for multicollinearity between the predictors. If multicollinearity exist, data transformation such as the earlier described Principal Component Analysis (PCA) has to be performed, otherwise the results of the paper can be questioned (32). Especially the assessment of individual characteristics is not possible if multicollinearity is present (33), which especially in this research, is problematic. The Variance Inflation Factor (VIF) between the predictor variables will be used as detection mechanism. According to Akinwande *et al.* (34), the acceptable values for VIF are between 1-5, greater than 5 could cause problems in the models. Fortunately, no VIF values greater than five are observed.

Although PCA is not necessary to remove multicollinearity from the predictors, it could still be useful for dimensionality reduction of the data. A scree plot is used to determine the number of components, Franklin *et al.* (35) used Kaiser's rule to determine how many components are retained. Kaiser's rule states that all components with eigenvalues larger than 1 should be kept, which is in our case 18.

The final two data modification techniques that will be implemented to test whether it enhances model performance is outlier detection and data augmentation. As the test data has mostly been recorded manually, there are some outliers that can be caused by personal errors. For data sets with a sample size larger than 80, an entry is considered if the standard score of an entry is ± 3 (36). However, considering outliers can also contain valuable information, models will be trained with and without outlier removal.

3.2.2 Event data sets

3.2.2.1 Feature engineering

For xSP and xSD estimation, a data set for each event that includes all the recorded occurrences with relevant features that are engineered using coordinate data at the timestamp of the event and (imputed) test results of the player performing the event is made. Based on the coordinate data, features of the events are derived such as pressure, pass angle, direction, speed, area on the pitch and pass length. The pressure feature is divided into several sub-features shown in the list below, which are all computed using the coordinates of the opponents when the event is performed.

- Pressure direction: the direction from where the player is pressured
- Distance to opponent: distance to the closest opponent to the player
- Mean distance: average distance to all opponents
- Distance back: the distance to the closest opponent behind the player
- Distance front: the distance to the closest opponent in front of the player
- Pressure level: no Pressure, limited Pressure or full Pressure, based on the other features

For xSD prediction different features are extracted out of the tracking data such as, distance, highest speed, average opponents within 1m/5m.

To evaluate a player's defensive contribution, a combination of the interception frequency and the defensive/offensive gain it provides will be used. The offensive gain is quantified using a feature called 'Expected threat', which expresses the probability that a shot on target happens within five passes from each position on the field. Figure 3.9 shows the values computed using shot events from the data. A red color indicates a high probability of a shot happening when starting of in that specific area. Unfortunately, there is not enough shot data to also evaluate a player's shooting performance in this study.

It is logical that areas closer to the goal of the opponent the probability of a shot happening within five passes increases. The corner spot also has a higher probability than the areas around it. What stands out is that the own five meter box has significantly higher probability than anywhere else in the own half. The reason for this is that a shot is more likely to happen within five passes as the goalkeeper tends to kick it long from that area.

Before features based on the position on the field are added, the position of the away team has to be inverted. This is done to ensure that the relative position of the pass is

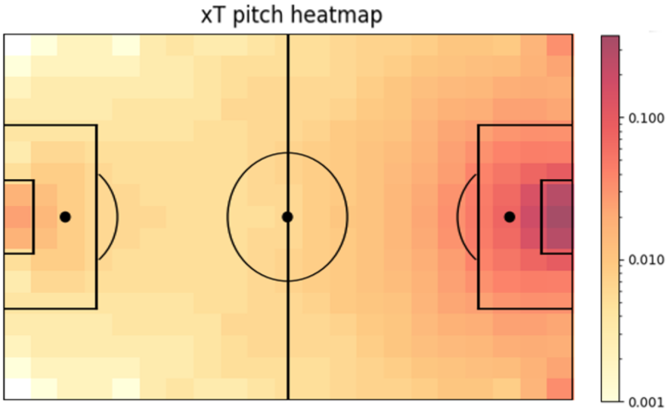

Figure 3.9: Expected threat on all areas on the pitch

the same for both teams, a pass at the back for the home team is on the other side of the pitch as a pass at the back for the away team. To get an indication the importance of the position of the pass Figure 3.10 indicates the heatmap of pass percentage based on position on the field and Figure 3.11 the heatmap with inverted coordinates for the away team. The legend of the heatmaps represents the color scale used, with the shades of green indicating the pass completion percentage. As expected, it is observed that the coordinates of the pass heavily influence the completion percentage, a higher percentage of passes is completed in defense than attack. But this is only the case when the coordinates are adjusted, without adjusting the percentages even out as passes in the defense and attack are on opposite sides of the pitch for both teams.

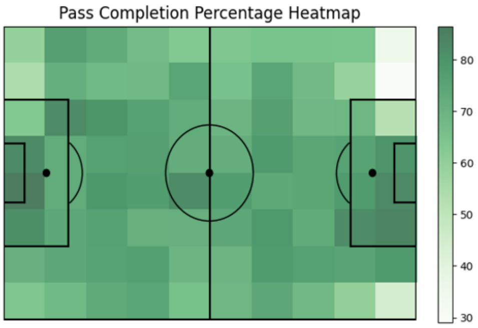

Figure 3.10: Pass completion percentage heatmap

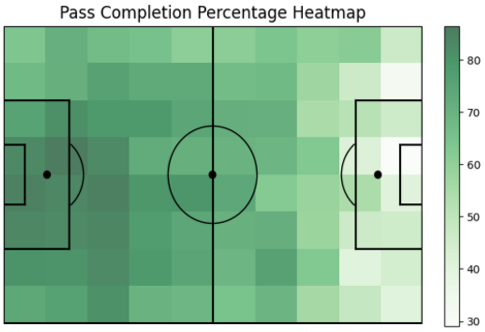

Figure 3.11: Pass completion percentage heatmap inverted for away team

It is also observed that 60.4% of the passes that are played forward are successful, while passes played sideways or backwards have a success rate of 76.4%. The contextual features of the pass have, as expected, correlations with the success rate. When plotting the distributions of test data for both successful and unsuccessful passes, it is observed that the distributions are identical. Similarities were expected, but identical distributions indicate a weak correlation between test data and pass completion.

Then, interval information of the player at the time of the event is added such as time playing, heart rate, physical load, etc. This is done to ensure that the test data is the only thing that differs among the players and to prevent players that are tired from having a disadvantage compared to players who just entered the field.

3.2.2.2 Class imbalance

When exploring the class frequencies, it is observed that all event classes are imbalanced. This is expected as high-level football players usually have a higher pass/dribble completion than 50%. The passing data is moderately imbalanced with 17364 successful passes recorded and only 7109 unsuccessful passes, while the dribbling data is significantly more imbalanced with 12,329 successful dribbles and only 2107 unsuccessful dribbles. A dribble is detected every time a player touches the ball before passing/shooting, even a single touch before passing the ball counts as a successful dribble. This does not define a player's dribbling skills, the paper by Dick (12) had a similar problem and filtered out the noise. Now, only dribbles with over 3 meter distance covered are used to only capture actual dribbles resulting in 11,119 successful and 1970 unsuccessful dribbles respectively. Several oversampling methods can be used to address the class imbalance, Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) are the classic oversampling methods. According to Brandt and Lanzén (37), neither is significantly better than the other, however SMOTE has a slightly better performance than ADASYN. Thus, this paper uses SMOTE to address class imbalance. SMOTE generates minority class samples using linear interpolation preventing the overfitting problem of Random OverSampling (ROS) (38). There are 4873 recorded interceptions, but interceptions are always successful, meaning that oversampling is not applicable.

The baseline xSP and xSD models will be made using the data set that contains only contextual features, then the aim is to improve this model by adding test results of the player who performs the event as features.

4

Methodology

The objective of this paper is to identify key characteristics of football players using test data, which is done using two types of machine learning methods. Classification will be used to assess the difficulty of passes and dribbles (xSP and xSD), these values are used to create descriptive metrics (PPR/DPR) to be able to accurately evaluate the performance of the players. Regression models are deployed to predict overall performance of the players. If only the overall performance is predicted, valuable information on how characteristics influence how players perform different event types will be lost. To address this, the classification model that only used contextual features of the event will be improved by adding test data of the player performing the event. This provides a more profound understanding on how characteristics influence different types of events.

4.1 Classification and Regression models

During data exploration it was observed that no strong linear correlations exist between the predictor variables and the target variable. Consequently, regression models that assume correlations between predictors and the target will not be effective. This rules out simple/multiple linear regression, polynomial regression and ridge regression.

4.1.1 Random forest

Random forest is a suitable machine learning for the available data set, as it can be used to capture complex data patterns in high-dimensional data (39).

4.1.1.1 Theoretical framework

The random forest machine learning model, originally developed by Leo Breimann in 2001 (40), will be used for both classification and regression. In his paper, Breimann proposes an ensemble method that combines the results of multiple decision trees through voting (classification) or averaging (regression). An overview of this process is depicted in Figure 4.1.

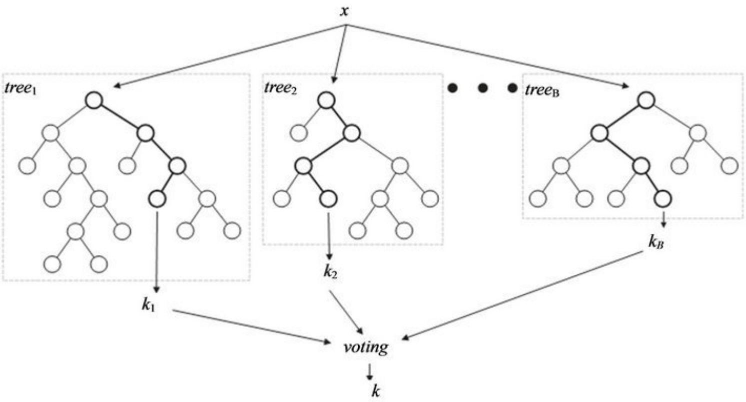

Figure 4.1: Illustration of the random Forest voting mechanism (41)

Breimann trains the individual decision trees using bagging; each tree is trained on a subset sampled with replacement of the training set. The diversity results in more robustness with respect to outliers and noise, additionally the risk of overfitting is reduced compared to fitting a single decision tree.

4.1.1.2 Practical implementation

The random forest algorithms are implemented in Python using the *sklearn* libraries *RandomForestClassifier* and *RandomForestRegressor* respectively. Tuning the models is also done using functions provided by *sklearn*. First, feature selection is performed using *SelectFromModel*, which transforms the training data such that only features with an importance greater than a threshold are kept. *SelectFromModel* is used as feature selection, using feature importances to capture complex non-linear relationships. Then, the hyperparameters are optimized using the *GridSearchCV* library, where hyperparameter configurations are exhaustively evaluated using k-fold cross-validation. The hyperparameter values that are tried are shown in the list below.

- n_estimators: 50, 100, 200, 500
- max_depth: None, 3, 5, 7
- max_features: auto, sqrt, log2
- min_samples_split: 2,5,10
- min_samples_leaf: 1,3,5,10

N_estimators defines the number of trees in the forest and *max_depth* defines the maximum depth of each of the trees in the forest. *Min_samples_split* and *max_features* define the minimum samples required for splitting an internal node and maximum number of features used to perform the split. Then $min\_samples\_leaf$ denotes how many samples should at least be present in a leaf node. Another parameter that will be carefully tuned is the oversampling ratio of SMOTE, as a standard ratio of 1:1 could skew the xSP predictions below the true values.

4.1.2 Gradient Boosting family

In this section, the general gradient boosting framework will be discussed, together with brief explanations of the algorithms belonging to the gradient boosting family that will be implemented. Gradient boosting algorithms will also be used for both classification and regression.

4.1.2.1 Theoretical framework

Similar to the random forest, gradient boosting is an ensemble method based on decision trees introduced by Friedman (42) in 2001. Contrary to the random forest, it does not independently create the trees on a subset of the training data but it sequentially adds trees to the model. Each tree that is added corrects the mistakes of the previous trees as they are trained on the residuals (43), finally a weighted combination of the individual trees are used to get a prediction. A schematic overview of this process is depicted in Figure 4.2

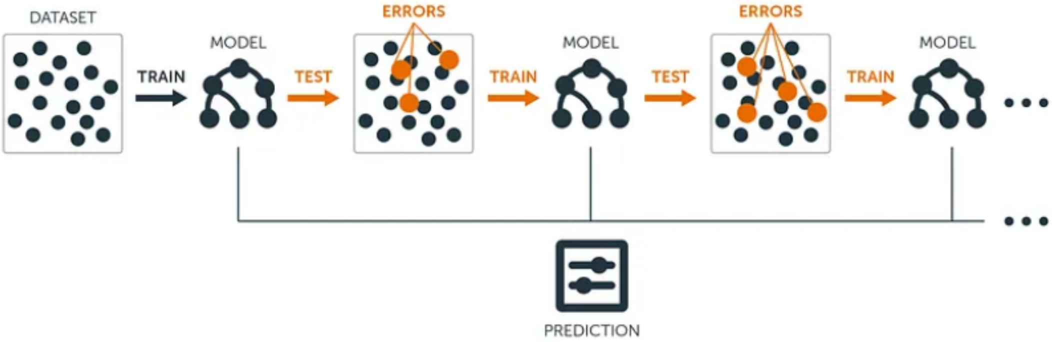

Figure 4.2: Gradient Boosting framework (44)

Friedman describes the Gradient Boost algorithm as 4 steps that are repeated m times which will be briefly explained. In this section, the equations used to show the calculations behind the steps are from his paper (42), where Gradient Boosting was first proposed. Before starting the loop, an approximation of the true function $F^*(x)$ has to be made, the initialization is given by Equation 4.1. Where $L(y_i, \rho)$ denotes the loss function that measures the discrepancy between the target y_i and the prediction of the model ρ.

$$F_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{N} L\left(y_i, \rho\right) \tag{4.1}$$

The first step is identify the samples that the model has predicted badly, which is done by calculating the pseudo-residuals $\tilde{y}i$. The pseudo-residuals show the magnitude of the error of each prediction and are calculated using Equation 4.2. The equation shows that the pseudo-residuals is the negative gradient of the loss function $L(y_iF(\mathbf{x}_i))$, with respect to the predictions of the ensemble $F(\mathbf{x}_i)$. With x_i and y_i denoting the input and output for the i-th sample.

$$\tilde{y}_i = -\left[\frac{\partial L\left(y_i F\left(\mathbf{x}_i\right)\right)}{\partial F\left(\mathbf{x}_i\right)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, N \tag{4.2}$$

As seen in Figure 4.2, these pseudo-residuals will be used to train a new decision tree. The mathematical representation of this process is given by Equation 4.3, where the optimal configuration of the m-th decision tree $\mathbf{a}_m$, is calculated by minimizing the loss function. The value $\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})$ represents the difference between the pseudo-residuals, $\tilde{y}_i$, and the predictions made by the new individual decision tree $h(\mathbf{x}_i; \mathbf{a})$ scaled with β and $\mathbf{a}$ defining a specific structure and the characteristics of the individual tree. By optimizing β and $\mathbf{a}$, the optimal configuration for the new decision tree is found.

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} \left[\tilde{y}_i - \beta h\left(\mathbf{x}_i; \mathbf{a}\right)\right]^2 \tag{4.3}$$

As discussed, the prediction of the ensemble is a weighted average of all weak learners (decision trees) in the ensemble. After creating a new weak learner, the next step is calculating the optimal weight it has in the final prediction. The weight is calculated using Equation 4.4, which is similar to the Equation 4.3, but now the predictions of the existing ensemble methods are taken into account to optimize the weight assigned to the new weak learner. The optimal weight ρ_m is calculated by minimizing the loss function, where the m-th tree multiplied with different values for ρ is added to the existing ensemble F_{m-1} based on the previous $m - 1$ decision trees.

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L\left(y_i, F_{m-1}\left(\mathbf{x}_i\right) + \rho h\left(\mathbf{x}_i; \mathbf{a}_m\right)\right) \tag{4.4}$$

Finally, the ensemble method is updated by adding the new learner defined using the earlier equations. Equation 4.5 demonstrates how the new ensemble is the old ensemble with the new learner weak added, weighted by the optimized weight p_m. This process of updating the ensemble model is repeated m times, or until convergence is reached.

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h\left(\mathbf{x}; \mathbf{a}_m\right) \tag{4.5}$$

The described steps are the same for all algorithms in the gradient boosting family, however some formulas might be slightly different. The XGBoost algorithm is a more advanced gradient boosting algorithm introduced by Chen and Guestrin (45). They add a regularization term and shrinkage to reduce the chance of overfitting, these techniques achieve this by reducing the complexity of individual trees and reducing the training step size. The modifications to the equations are the addition of the regularization term $\Omega(T)$ to the loss function and shrinkage adds new weights η after each iteration.

LightGBM, introduced by Ke *et al.* (46), also uses regularization and further improves the original gradient boosting algorithm by implementing Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to improve efficiency by using less data and bundle exclusive features.

The advantage of gradient boosting algorithms is that imputation is not a necessity, for LightGBM and CatBoost one-hot encoding of categorical features is also not necessary. However, the results of both inputs will be evaluated.

4.1.2.2 Practical implementation

XGBoost, LightGBM and CatBoost all have their respective libraries in python that implement the algorithms described above for both classification and regression. The hyperparameters for both regression and classification implementations are the same and will again be tuned using the *GridsearchCV* library. The *SelectFromModel* from the *sklearn* library will be used for feature selection. Considering all three are gradient boosting algorithms, the algorithms have similar parameters. The list below shows the values for the shared parameters of the three models.

- learning_rate: 0.001, 0.01, 0.05, 0.1
- n_estimators: 50, 100, 200, 500
- max_depth: none, 3, 5,8, 10
- subsample: 0.5, 0.7, 0.9
- reg_lambda (or l2_leaf_reg): 0.01, 0.05, 0.1

Considering both the random forest and gradient boosting models are tree based ensemble methods, it is no surprise that several parameters are the same. *Subsample* defines the ratio of samples from the training data that is used. The *learning_rate* and *reg_lambda* are both parameters that prevent overfitting in the model. A lower learning rate reduces the step size during the learning process, giving each weak learner less influence in the final result, resulting in a slower convergence. *Reg_lambda* (l2 leaf reg for CatBoost) which is L2 regularization that adds a penalty to the loss function or to the leaf weights. XGBoost and LightGBM also have L1 regularization with the parameter *reg_alpha*. LightGBM also has *num_leaves* as parameter which determines the maximum number of leaves the tree has, values between 10-30 are added to the gridsearch.

4.1.3 K-Nearest Neighbors (KNN)

Despite it being introduced by Fix and Hodges (47) in 1951 and considered an old algorithm, the K-Nearest Neighbors algorithm is still effective and often used nowadays. As discussed with the KNN imputer, it is non-parametric (48) and is able to handle complex, non-linear relationships (49) making it a suitable algorithm for this research.

4.1.3.1 Theoretical framework

KNN is based on the similarity of the samples in the data set. As indicated by the name, it classifies a sample based on its k nearest neighbors, where k is a predefined number that represents the number of neighbors used in the prediction. The distance between samples is calculated with a distance metric, often the Euclidean distance or Manhattan distance. These two distance metrics can be generalized using the Minkowski distance (50) of which the formula is shown in Equation 4.6.

$$d_{Md}\left(X_i, X_j\right) = \left(\sum_{t=1}^{m} \left|x_t^i - x_t^j\right|^p\right)^{1/p} \quad \text{for } p \geq 1 \tag{4.6}$$

Taking $p = 1$ results in the formula for the Manhattan distance and the formula is equivalent to the Euclidean distance when $p = 2$. Increasing p further would result in higher sensitivity towards larger values.

4.1.3.2 Practical implementation

Sklearn provides the libraries $KNeighborsClassifier$ and $KNeighborsRegressor$ which will both be implemented. However, feature selection cannot be done using the $SelectFromModel$ library for KNN as it is a distance-based algorithm. *Sklearn* does provide another feature selection method called $SelectKBest$ that selects features based on individual correlations with the target. As mentioned above, the optimal value of k is subject to controversy whether to use a low value to prevent overfitting or a higher value such as $\sqrt{k}$. The fact that it is an older algorithm is observed in its simplicity, only a few parameters have to be tuned. The configurations are depicted in the list below.

- n_neighbors: 3, 5, 10, 20, 30, 40
- weights: uniform, distance
- p: 1,2

Due to the uncertainty of the suitable value for k, six values will be tested for $n\_neighbors$. The *weights* parameter defines the weight of each neighbor, uniform assigns the same weight to all k neighbors while distance assigns higher weights to closer neighbors. As explained, p defines whether the Manhattan distance or Euclidean distance will be used as distance metric.

4.1.4 Artificial Neural Network

The Artificial Neural Network (ANN) is inspired by the human brain and uses a highly interconnected network of neurons that mimic the processes of real neurons (51). As our data is of tabular form without temporal nature, the best suited architecture type is a Multi-Layer Perceptron (MLP), which is a feed-forward neural network.

4.1.4.1 Theoretical framework

It has been shown that the MLP is able to approximate every measurable function, even highly non-linear, without making assumptions about the underlying distribution of the data(52) (53). The Multi-Layer Perceptron (MLP) consists of connected neurons in an input layer, one or more hidden layers and an output layer. The basic structure of a MLP is depicted in Figure 4.3.

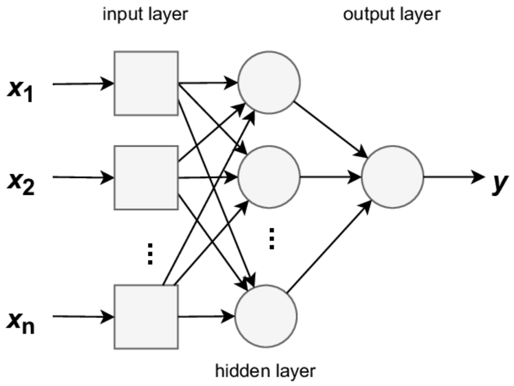

Figure 4.3: Structure of MLP (54)

The value of feature i, illustrated as x_i in the figure, is propagated through the input layer to the neurons in the hidden layer. Each neuron in the hidden layers computes a weighted sum of values it receives and adds a bias, denoted as $\sum_{j=1}^{n}(x_j w_j) + b$. Before propagating it to the next layer, an activation function σ is applied to introduce non-linearity. The most common activation functions are the sigmoid, reLU, tanh or the softmax function. Figure 4.4 shows this process graphically, together with the mathematical formulation of the activation function being applied to the weighted sum.

The weights will be initialized randomly and then optimized using backpropagation which, like gradient boosting, consists of several steps that will be iteratively repeated. Backpropagation was originally introduced in 1975 but began to gain popularity after a

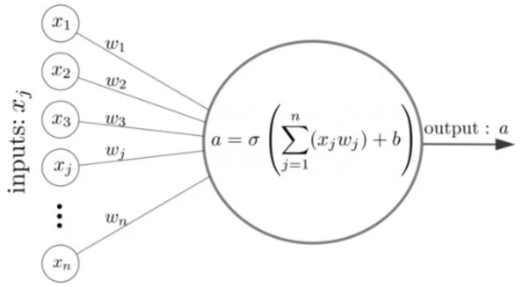

Figure 4.4: Process in a single neuron of a MLP (55)

paper by Rumelhart *et al.* (56) demonstrated the effectiveness of the algorithm. To begin, a training sample will be forward propagated through the network with the process explained above. Then, the error of the output with respect to the target value is calculated using a loss function. The gradient of the error is computed and backpropagated through the network by computing the gradients of the error with respect to each weight one layer at a time. Equation 4.7 represents the gradient of the error with respect to the weighted sum z^l_j at node j of the MLP. The partial derivative of the error with respect to the activation $\frac{\partial E}{\partial a^l_j}$, multiplied by the partial derivative of the activation with respect to z^l_j, $\frac{\partial a^l_j}{\partial z^l_j}$:.

$$\delta^l_j = \frac{\partial E}{\partial a^l_j}\frac{\partial a^l_j}{\partial z^l_j} \tag{4.7}$$

The resulting δ^l_j, quantifying the contribution of the activation of neuron j in layer l to the error of the overall network, can be used to calculate the gradient of the error with respect to the weights using the chain rule, see Equation 4.8. δ^l_j is multiplied with the activation a^{l-1}_i of the neuron in the previous layer. The gradient of the error with respect to each weight is used in computations for the weight updates.

$$\frac{\partial E}{\partial w^l_{i,j}} = a^{l-1}_i \delta^l_j \tag{4.8}$$

After this, an optimizer is used to iteratively update the weights and biases in the network such that the loss function is minimized. This process is repeated until convergence by approaching the minimum of the loss function or when the specified number of epochs is reached. Gradient descent is the most basic optimizer which updates the weights by subtracting the original weight by the gradient of the total loss function with respect to W_j multiplied with the learning rate as seen in Equation 4.9. The total loss function of weight w_j is the sum of the gradients of the loss function with respect to w_j over all m samples. The learning rate determines the step size each iteration and tuning it correctly is critical. A learning rate that is too high could fail to converge and if it is too small it takes too long to converge.

$$\omega_j^* \leftarrow \omega_j - \alpha * \nabla_w \sum_1^m L_m(w) \cdot \hat{w}_j \tag{4.9}$$

Gradient descent has several disadvantages, the computational efficiency is low as it uses all training samples each batch. Even though this results in smooth convergence, the lack of randomness means that it can converge to a sub-optimal solution and thus can get stuck in a local minima on non-convex functions(57).

By introducing noise and randomness, stochastic gradient descent or the Adam optimizer reduce the risk of getting stuck in a local minima as opposed to the batch gradient descent optimizer. Stochastic gradient descent only uses the weight update formula one sample each batch which is observed in Equation 4.10 as the summation is removed over all m samples. Mini-batch gradient descent is in-between both earlier mentioned optimizers, it uses a small number of samples each batch.

$$\omega \leftarrow \omega - \alpha * \nabla_w L_m(w) \tag{4.10}$$

Figure 4.5 shows how stochastic gradient descent converges less smooth due to the introduced randomness. As discussed, this can help escape local minima.

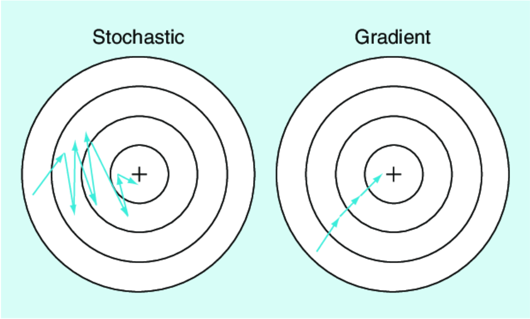

Figure 4.5: Convergence comparison of stochastic/batch gradient descent (58)

Since the configuration of the learning rate has a significant impact on the results, a big advantage of the Adam optimizer is the iterative adjustments of the learning rate during training, making it a more robust optimizer. The formula of the Adam optimizer is given by Equation 4.11, where $\hat{m}_t$ and $\hat{v}_t$ are estimates of the first and second moment of the gradients respectively. These are used to adapt the learning rate for each individual weight. ϵ is a small value that prevents a division by zero.

$$w_{i,j}^* \leftarrow w_{i,j} - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{4.11}$$

The difference between MLP in regression and classification is the loss function used, the most common loss function for regression is the Mean Squared Error (MSE) and the loglikelihood for classification. Even though a training set of 500 players is relatively small for deep learning methods, a MLP will be made for both classification and regression.

4.1.4.2 Practical implementation

The MLP is implemented in python using the keras modules, which are a part of the $tensorflow$ package. As both data sets used in this paper are not excessively large, neural networks of one or two hidden layers will be used to prevent the model from overfitting. In addition to different architectures, various optimizer configurations and the four activation functions mentioned will be tested. Finally, the learning rate and number of epochs will be tuned. The values tested for the learning rate are similar to the previous models; 0.001, 0.01, 0.05, 0.1, while the number of epochs is optimized by plotting the validation and test loss. The optimal number of epochs is approximately at the point when the training loss decreases and the validation loss starts to increase or remains stable, indicating an optimal trade-off between training performance and generalizability.

4.2 Evaluation metrics

4.2.1 Classification

For classification, the most used metrics are the accuracy, recall, precision and F1-score derived from the confusion matrix. However, as previously discussed, these metrics fail to show how well a model is calibrated. A model that is able to estimate the xSP accurately is preferred over a model that predicts more passes correctly as the purpose of the model is to estimate the difficulty of passes. The issue is that the true xSP of a pass is unknown. To overcome this issue, the *empirical Brier* is used that approximates the Brier score by using the sample outcomes. The empirical Brier score defines the sharpness and calibration of the model and is calculated using Equation 2.1. A calibration plot is used to visualize the calibration of the model across several ranges of probabilities. The predictions of the model are split into bins, the model is well calibrated model if in each bin the proportion of positive values is equal to the average xSP. The result is a scatter plot with the average predicted xSP against the true proportion of successful passes of each bin. An overconfident model would predict higher values than the true proportion which would be observed if the point in the scatter plot is above x=y and vice versa.

While the Brier score and a calibration plot are often used as metrics to evaluate the calibration of a model, both are not a quantitative measure of the calibration. The calibration plot shows a general visualization of the calibration, while the Brier score assesses both the calibration and the sharpness of the model (59). To formally assess binary predictions, D.J. Spiegelhalter (60) decomposed the Brier score to the Spiegelhalter's z-statistic, the formula is given by Equation 4.12 and it uses the standard normal distribution to measure whether a model is well calibrated (61). With x_i and p_i denoting the true label and predicted xSP of i. Similar to the calibration plot it compares observed and predicted frequencies, a higher value indicates a better calibration. The null hypothesis of the statistical test is that the model is well calibrated, the model is poorly calibrated if the resulting z-score is outside the range of -1.96-1.96 and thus significant using the conventional $\alpha = 0.05$ (62) (63).

$$Z(p, x) = \frac{\sum_{i=1}^{n} (x_i - p_i)(1 - 2p_i)}{\sqrt{\sum_{i=1}^{n} (1 - 2p_i)^2 * p_i * (1 - p_i)}} \tag{4.12}$$

The final metric used is the Expected Calibration Error (ECE) which measures the weighted average of all differences between the predicted and actual fraction in each bin. But since no single metric gives a complete assessment of a model's performance, a combination of these metrics are applied for a comprehensive overview of the model's performance.

4.2.2 Regression

For regression, the Root Mean Squared Error (RMSE) and Mean Average Error (MAE) are used to measure the magnitude of the errors and the R^2 is used to measure the goodness of fit of the model performance. The R^2 represents how much the variance in the target variable can be explained by the predictor variables. The main goal of the research is explaining performance using test results, which means that the main focus is on the R^2. RMSE and MAE can provide additional numerical evaluation metrics to compare models if the R^2 values are similar. In addition to these metrics, a scatter plot of the predictions versus the true values is made to identify a possible pattern of prediction errors.

4.2.3 Player performance

While the optimal xSP and xSD models are found using the metrics described above, the performances of the players also have to be evaluated using fair metrics. A player is evaluated based on his Passing Performance Ratio (PPR), Dribbling Performance Ratio (DPR) and interception rating values. The PPR and DPR are computed using the predicted xSP and xSD values by the optimal classification models. The PPR of a player is calculated by dividing his number of successful passes by the number of successful passes the xSP model predicted him to give. The predicted number of successful passes for a player is simply the sum of the xSP values of all passes that the player gives. The mathematical formulation is shown in Equation 4.13.

$$PPR_x = \frac{\sum_{i \in P_x} y_i}{\sum_{i \in P_x} xSP_i} \tag{4.13}$$

The equation shows the calculation of the PPR for player x, denoted as PPR_x, where P_x is the set of all passes given by player x. y_i is a binary value equal to 1 if pass i is successful and 0 if pass i is unsuccessful. xSP_i denotes the probability that pass i is successful according to the optimal xSP model. The PPR of player x is then computed by summing the y_i of all passes made by player x ($\sum_{i \in P} y_i$) and dividing it by the summation of the success probabilities xSP_i of all passes given by player x (denoted as $\sum_{i \in P} xSP_i$). A PPR greater than 1, signifies that the player outperforms the expectations of the model, while a value lower than 1 indicates a performance lower than expected. Unlike pass completion percentage, the PPR does not penalize players that perform harder passes, to illustrate this a small example is given. Consider player A that gives 10 successful passes out of 12 with an average xSP of 0.95, his PPR is $\frac{10}{12*0.95} = 0.87$ and a pass completion percentage of 83%. If Player B gives 8 successful passes out of 12 with an average xSP of 0.57, he has a PPR of $\frac{8}{12*0.57} = 1.17$ with a completion percentage of 66.7%. So, even though Player B has a lower completion percentage, he has a better PPR as he delivers significantly more difficult passes than Player A.

The calculation of dribble performance (DPR) is identical and shown in Equation 4.14.

$$DPR_x = \frac{\sum_{i \in D_x} y_i}{\sum_{i \in D_x} xSD_i} \tag{4.14}$$

Finally, the interception rating is a weighted average of the scaled values of the frequency, defensive contribution and offensive gain of a player's interceptions. The frequency is defined as the number of interceptions a player performs per minute. The offensive gain is based on the location of the interception quantified using the expected threat values for each position on the pitch defined in Chapter 3. Lastly, the defensive contribution is defined as the number of teammates that are in front of the ball (TFB). This means that an interception becomes more important when only a few teammates are left to defend the ball as opposed to when the whole team is still able to defend. To ensure a balanced contribution of each of the three factors, the three values are scaled before being used in the formula shown by Equation 4.15.

$$IR_x = (0.6 * freq_x) + (0.2 * \frac{\sum_{i \in I_x} xT_i}{|I_x|}) + (0.2 * \frac{\sum_{i \in I_x} TFB_i}{|I_x|}) \tag{4.15}$$

The formula shows how the interception rating of player x (IR_x) is calculated. First, the player's interception frequency x ($freq_x$) is multiplied by 0.6. 0.6 is chosen as the weight because the number of ball recoveries is deemed to be the most important of the three factors. The set of interceptions performed by player x is denoted by I_x, $\frac{\sum_{i \in xT_i}}{|I_x|}$ is the average offensive gain of the interceptions in I_x. Then the average defensive contribution is computed using $\frac{\sum_{i \in TFB_i}}{|I_x|}$. Note that, as discussed, the $freq_x$, xT_i and TFB_i are scaled.

5

Results and Evaluation

This chapter will present and analyze the results of both the models that estimate xSP and xSD, as well as the models that predict the overall performance of the players. For the best performing models, the metric values together with the selected features and hyperparameter configurations will be discussed thoroughly. In Section 5.1, the xSP and xSD are estimated to create the PPR and DPR metrics, then the results of the overall performance prediction will be shown in Section 5.2. To avoid repetitiveness, only relevant plots are shown in this section; less relevant plots are omitted from the report or can be found in the appendix.

5.1 Event classification

In this section, the difficulty of passes and dribbles will be estimated. First, the models only using contextual features will be evaluated, then test data will be added to see whether a more comprehensive understanding of the influence of the test results on specific event types can be achieved. Finally, the resulting probability predictions will be validated using common sense and general knowledge of the sport.

5.1.1 xSP

Table 5.1 shows the metric scores for all xSP models only using pass features, to establish the metric that also evaluates pass difficulty. The table shows that all models are able to accurately estimate the xSP, but the random forest and XGBoost models exhibit the best overall performance on the three metrics. Fairchild (6) improved its referenced xG model by obtaining a Brier score of 0.19, all our models achieve even lower values than 0.19 indicating well calibrated models. All models have high Spiegelhalters p-values, indicating that no model is poorly calibrated. The average difference between the predicted probabilities and the observed frequencies within each bin is also small, being approximately 0.02. The models are obtained using scaling, KNN imputation and without using SMOTE oversampling and PCA. Using SMOTE or PCA disrupted the distribution of the passes and higher probabilities were predicted for lower probability passes and lower probabilities for the higher probability pass bins. The calibration plots using SMOTE 1:1 oversampling and PCA can be seen in Figure A.1 and Figure A.2 in the Appendix.

Model	**Brier score**	**Spiegelhalters p-value**	**ECE**	**#features**
Random Forest	0.171	0.739	0.021	16
XGBoost	0.168	0.722	0.024	15
LightGBM	0.171	0.713	0.026	12
CatBoost	0.172	0.688	0.038	10
KNN	0.180	0.642	0.020	20
MLP	0.178	0.688	0.028	-

Table 5.1: Evaluation xSP models excluding test features

As the metric values for the Random Forest and XGBoost are very similar, their calibration plots will be compared as well to decide the superior model. The calibration plots of the random forest and XGBoost are depicted in Figures 5.1 and 5.2. It is observed that there is a very close alignment between the average predicted xSP in a bin and the fraction of successful passes in a bin for both models. However, the random forest seems to be slightly better calibrated for passes with lower success rates. As the linear fit of the random forest is slightly better, it is the best model for estimating the xSP.

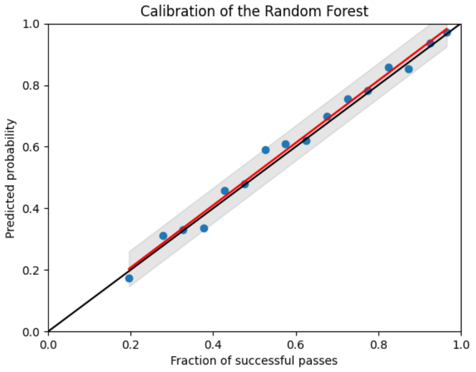

Figure 5.1: Calibration plot RF with pass features

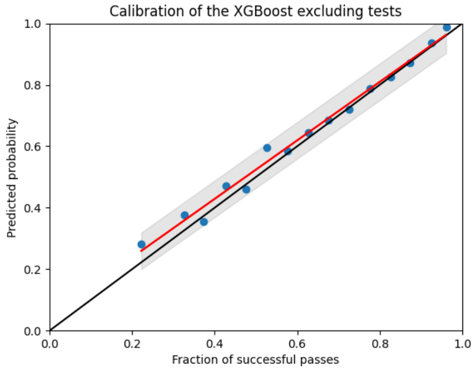

Figure 5.2: Calibration plot XGBoost with pass features

The random forest uses the following parameter configuration: $n\_estimators : 500, max\_depth : None, max\_features : auto, min\_samples\_split : 2$, $min\_samples\_leaf : 2$.
The large number of estimators and no depth maximum, enables the random forest to capture complex patters. The automatic feature selection mitigates the risk of overfitting as it helps the model to only focus on important features. Finally the minimum samples requirements prevents the model from creating specific rules making the results more generalizable. In combination with these parameters, sixteen features are used by the model which are ranked by importance in Figure 5.3. The most important features are the pass length, angle, position on the pitch and several pressure features which seem reasonable.

Figure 5.4 shows for each player the expected successful passes based on the cumulative xSP against the true number of successful passes the player has given. The figure further confirms that the xSP is accurately estimated, as the points are evenly distributed around the x=y line, indicating the absence of systematic bias or over-/underestimation.

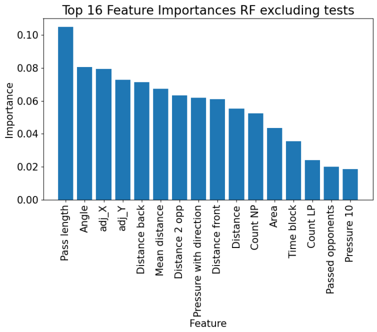

Figure 5.3: Feature importances RF excluding test data

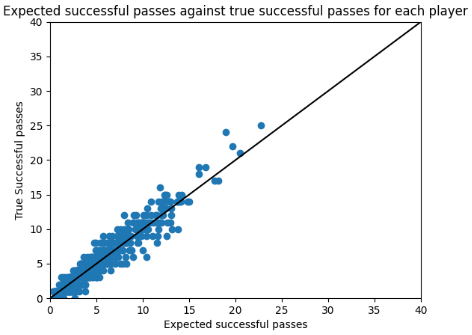

Figure 5.4: Expected successful passes against true successful passes for each player

The predicted xSP for each pass by the baseline model is used to calculate the PPR for each player. By integrating data about the passer, the probability estimation of a pass being successful can be improved and give valuable insights into the integral skills of a player for specific types of passes. Table 5.2 shows that all models except LightGBM decrease in performance.

Model	**Brier score**	**Spiegelhalters p-value**	**ECE**	**# features**
Random Forest	0.178	0.697	0.038	30
XGBoost	0.178	0.713	0.027	28
LightGBM	0.176	0.705	0.023	50
CatBoost	0.180	0.588	0.045	27
KNN	0.187	0.711	0.025	19
MLP	0.195	0.603	0.035	-

Table 5.2: Evaluation xSP models including test features

Consequently, the LightGBM model performs best, the calibration plot and feature importances are depicted in Figures 5.5 and 5.6. The calibration remains good, but is worse than only using pass features. The feature importances show that the first test score is the 21st most important feature, indicating the insignificance of the test features. For the other models, the best performance was obtained using fewer features, but a constraint was implemented that at least three test scores must be used to observe the effect it has. So for single passes, adding player skill sets as features only adds noise. However, it could still be that the test results are effective in the prediction of overall performance. Considering the predicted probabilities are less reliable than the baseline model, no insights are obtained on the effect of test data on specific pass types.

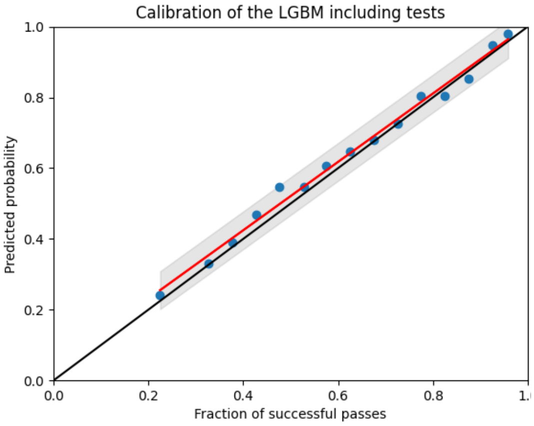

Figure 5.5: Calibration plot LightGBM with test features

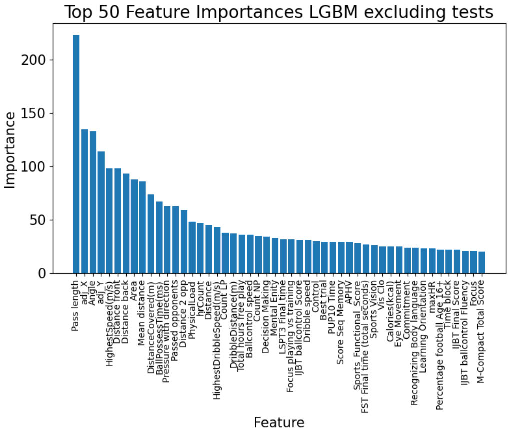

Figure 5.6: Feature importances LightGBM with test data

5.1.2 xSD

The same models can be applied to the dribble events which, as seen during the data exploration, was significantly more imbalanced. For xSP estimation, using SMOTE oversampling decreased performance, however, for xSD estimation oversampling to a 0.25/0.75 instead of the original class balance improved performance. Without oversampling, the bin sizes of the harder dribbles were too small and a noisy calibration was observed. The imbalance also meant that the majority of the xSD predictions are over 0.7 and that the brier score is heavily influenced by this majority class. The calibration plots of the models showed an accurate estimation of higher xSD dribbles, but a few struggled with the lower xSD dribbles, which was not reflected in the Brier score because of the dominance of the well estimated majority class. As seen in Table 5.3, the random forest and catboost demonstrate the best performance, with the MLP being slightly worse.

Model	Brier score	Spiegelhalters p-value	ECE	# features
Random Forest	0.111	0.677	0.050	22
XGBoost	0.128	0.421	0.063	10
LightGBM	0.121	0.533	0.067	24
CatBoost	0.109	0.712	0.044	10
KNN	0.118	0.349	0.070	10
MLP	0.111	0.640	0.059	-

Table 5.3: Evaluation xSD models excluding test features

The performance of catboost seems to be superior; however, the calibration plots of both the random forest and catboost will be compared to confirm this, as both have very similar results. Figures 5.7 and 5.8 show the respective calibration plots, apart from one bin, the catboost indeed shows a better calibration than the random forest. Notably, the catboost model was the worst model for xSP estimation.

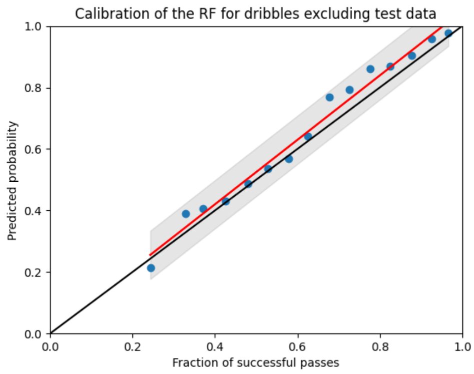

Figure 5.7: Calibration plot random forest for dribbles

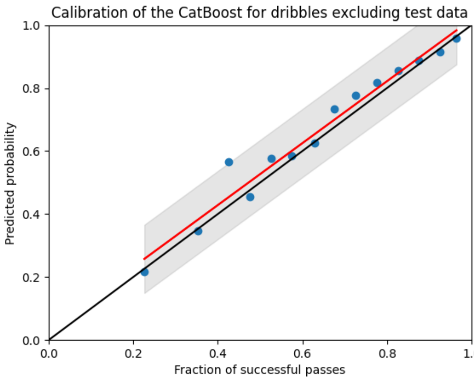

Figure 5.8: Calibration plot catboost for dribbles

The catboost model used ten features which are seen in Figure 5.9 with an optimal parameter configuration of $colsample\_bylevel : 0.5, min\_child\_samples : 2, n\_estimators : 100, subsample : 0.5$. The parameter configuration shows a balance between randomness and regularization. It is interesting that the number of teammates in front of the ball has a higher predictive power than the area on the pitch, suggesting that context is more important than position on the pitch. The pressure on the dribble, here defined as 'opponents in 1m radius', and the increase in xT are also strong indicators of the difficulty of a dribble. Difficult dribbles also tend to have a higher max speed as players often accelerate to blow by opponents.

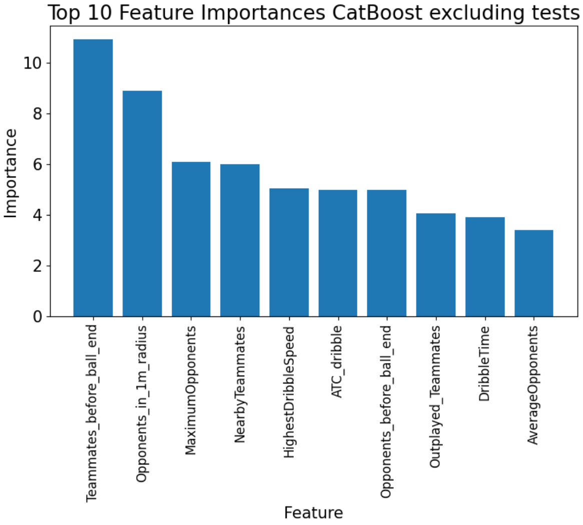

Figure 5.9: Ten features used by the optimal catboost model to predict xSD

The results of the models after the test data is added is not shown, as it had the exact same results as the models without. Contrary to the xSP models, where the test data had a small feature importance, the test data had for almost all models a feature importance of smaller than 0.01.

5.1.3 Result validation

The plots and metrics show that the xSP and xSD are well estimated; however, it is important to verify the numbers visually to judge whether the probabilities make sense using football knowledge. Figure 5.10 shows the situation of a hard pass, the green dots are the players that pass/receive the blue dot which represents the ball, while the red dots represent the opposing team (goalkeeper not shown). The situation resembles a defender/goalkeeper trying to pass to a forward using a long ball with slight pressure from the right. The model predicted a 30.5% chance the pass is successful, which seems reasonable as the striker could win the header but defenders are generally more likely to win aerial duels. According to FBRef (64), the sixteen players with the best aerial duel win percentage are defenders, with the highest ranked attacker being placed 55th in the Premier Leagues 2022-2023 season. Figure 5.11 shows a pass backwards with 97.8% success probability according to the model, which seems justifiable as it appears to be a simple backwards pass.

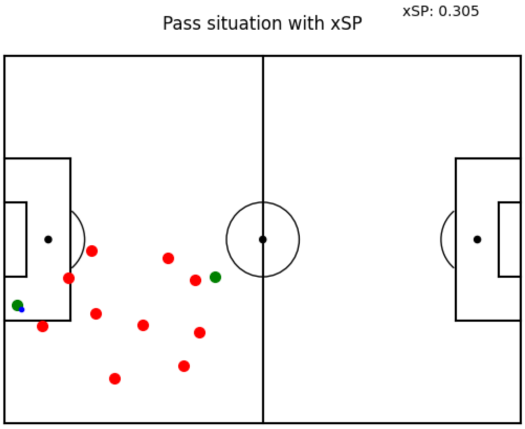

Figure 5.10: Pass situation of a hard pass according to the model

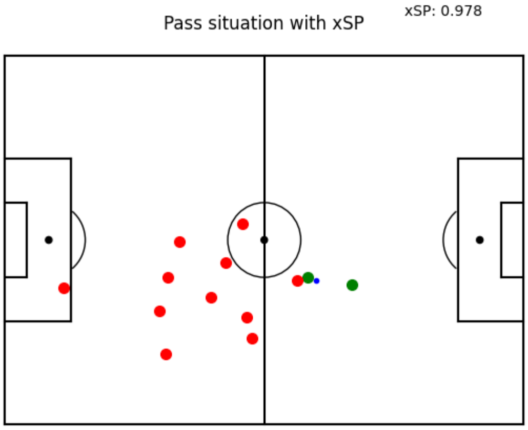

Figure 5.11: Pass situation of an easy pass according to the model

Dribbles are harder to visualize as it cannot be depicted in a single situation overview plot like the ones used for illustrating passes. Based on some of the key features from Figure 5.9, it was decided to give a general overview of the dribble using two position features, two pressure features and two dribble features. Table 5.4 shows two examples for easy, medium and hard dribble difficulties.

The two hardest dribbles are characterized by the player being the farthest up the pitch of their team, which was also the feature with the highest predictive power. Most papers

xSD	Teammates before ball	ATC	Opp 1m radius	Avg opp within 5m	Time	Highest speed (m/s)
0.119	1	-0.01	2	2.649	15.2	3.162
0.131	0	0.06	1	1	5	7
0.514	2	0.001	1	0.884	8.4	2.236
0.637	4	0	1	2.260	4.4	2.0
0.894	4	0.002	1	0.862	5.6	3.605
0.961	7	-0.004	0	0	2.8	4.242

Table 5.4: Dribble characteristics and their corresponding xSD

used attacking contribution as main feature, but the ATC alone is not enough to distinguish dribbles on difficulty. The hardest dribble even has a negative ATC, the model has predicted an 11.9 percent chance of completing it due to the high pressure and long dribble time. Dribbles predicted around 50% are a little more in the middle of the field with either being a longer dribble or slightly pressured. Finally, the easiest dribbles are short dribbles in defense. Overall, according to the calibration plot and rational reasoning using football knowledge the estimated xSD is accurate.

5.2 Overall performance

In this section, the overall performance of the players will be predicted using their test results. Assessing a player's performance should be done using a metric that is independent of the test data, the baseline xSP and xSD estimations are used for the PPR and DPR calculations.

This section is split into three subsections; the PPR will be predicted in Subsection 5.2.1, the DPR in Subsection 5.2.2 and finally, the interception rating in Subsection 5.2.3

5.2.1 Passing performance

The PPR of a player will be calculated using the estimated xSP values and the preprocessed test dataset. Table 5.5 shows the performance of the models on the RMSE, MAE and R^2 metrics. It immediately stands out that the R^2 values of the models are very low, this does not mean that the models are useless. Although the entire disparity in PPR can not be explained using test results, valuable information can still be extracted from the results as some correlation is observed, providing insights into important characteristics/skills. Accurately predicting the PPR of a single player is not the final purpose of the model, the goal is to find correlations between test results and performance, any correlation can provide information.

As discussed, our primary focus is on the R^2 metric in which XGBoost performs significantly better than the other implemented models. The parameter configuration of the XGBoost model used to obtain this result is: $learning\_rate : 0.1, n\_estimators : 50, max\_depth : 8, subsample : 0.5, reg\_lambda : 0.001, reg\_alpha : 0.001$. Together with the fifteen features illustrated together with their respective importance in Figure

Model	RMSE	MAE	R^2	# Features
Random Forest	0.087	0.065	0.071	24
XGBoost	0.091	0.072	0.175	15
LightGBM	0.092	0.070	0.052	14
CatBoost	0.099	0.85	0.033	18
KNN	0.084	0.064	0.053	18
MLP	0.094	0.074	0.044	-

Table 5.5: PPR prediction performance of the models

5.13. Figure 5.12 shows the true PPRs against the predicted PPR values, where each blue dot represents a player. As indicated by the low R^2 value, the model is not able to accurately estimate the PPRs of the players. However, the observed linear fit is upwards, indicating a pattern, albeit a small one. Because the entire disparity could not be explained using the test results, players who excel at the most important skills do not always outperform players who do not. However, they are more likely to perform well.

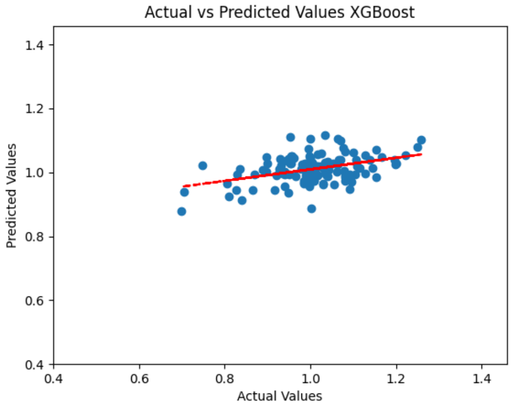

Figure 5.12: Scatter plot of XGBoost PPR predictions against true PPR

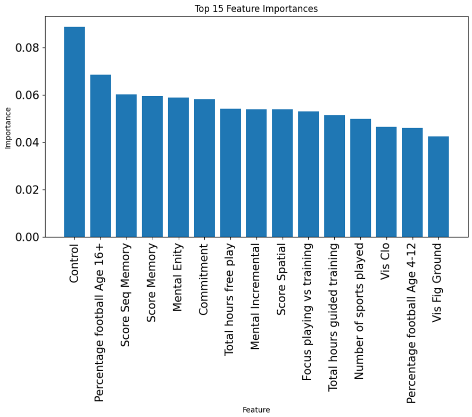

Figure 5.13: Feature importances XGBoost for PPR prediction

Notably, no features resulting from the PUP10 test are present in the top fifteen features used to predict the PPR. The most important characteristics and skills to have according to the model is control, describing how a player deals with his emotions. Next to controlling emotions, the two mental attitude features hold great significance. Then, the devotion of the player to the sport player when turning sixteen has the highest importance, this further validates the research conducted by Cote *et al.* (27). The importance of devotion is further confirmed by the importance of commitment and number of hours of free play as they also serve as indicators for a player's passion and enthusiasm for the sport. Interestingly, four of the memory tests are included in the fifteen features used, indicating that players with a quick and good memory tend to remember the positions of their teammates, reducing time needed for visual exploration.

To summarize, characteristics that display the highest importance in the prediction of the passing ability of youth players are the mental strength/attitude, devotion to the sport and their memory, all being cognitive abilities. To see whether a youth player is a good passer, just conducted a passing test, such as the PUP10, does not indicate whether the player passes well in matches. The fifteen features used to predict performance are predominantly mental and cognitive features rather than physical tests.

5.2.2 Dribble performance

Similarly to the PPR, the DPRs will be calculated using the resulting xSD values. The results are presented in Table 5.6, the R^2 values indicate again that the full performance cannot be explained using test results but correlation is observed.

Model	**RMSE**	**MAE**	**R^2**	**# Features**
Random Forest	0.109	0.083	0.097	17
XGBoost	0.154	0.109	0.066	12
LightGBM	0.129	0.098	0.106	19
CatBoost	0.131	0.100	0.072	20
KNN	0.114	0.096	0.114	20
MLP	0.134	0.096	0.038	-

Table 5.6: DPR prediction performance of the models

Even though the KNN model was one of the worst models for passing, it exhibits the best performance for dribbling prediction. Figure 5.14 shows the predicted DPR against the true DPR, a (small) upwards trend is observed. The twenty selected features are shown in Figure 5.15.

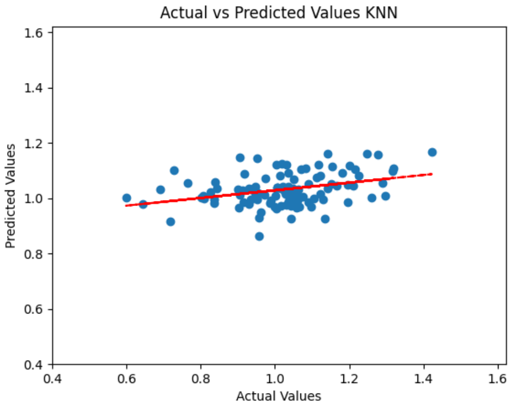

Figure 5.14: Scatter plot of DPR predictions against true PPR of KNN

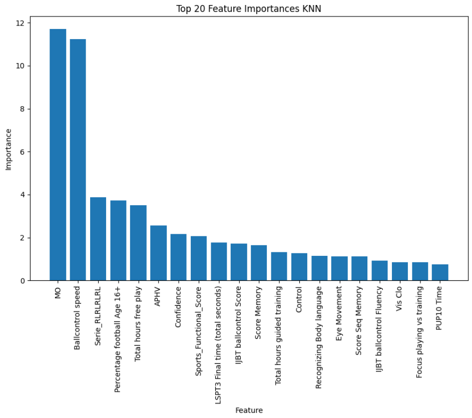

Figure 5.15: Feature importances KNN for DPR prediction

There are two features that are significantly more important than the others. While for passing the most important tests weren't tests related to passes, the two most important features for dribbling are Maturity Offset and ball control speed. MO is the number of years before a player reaches his/her age of peak height velocity, even though it is not necessarily a skill it is logical that this feature holds such significance. Players who experience an early growth spurt and physically more developed are anticipated to perform better dribbles. Indicating that if a player doesn't show extremely good dribbling performance but also hasn't already reached its APHV, it is too premature to their dismiss potential. Exceptional dribbling performance before APHV even signifies high potential. The relevance of the ball control speed feature speaks for itself. So contrary to the passing, the two main features used to predict performance are physical skills and characteristics.

5.2.3 Interception performance

Finally, the ability to frequently perform useful interceptions will be predicted using each players respective test results. Table 5.7 shows the model results, what stands out is that the results are significantly worse than for passing and dribbling.

Model	**RMSE**	**MAE**	**R^2**	**# Features**
Random Forest	0.566	0.492	0.007	24
XGBoost	0.565	0.423	0.012	20
LightGBM	0.555	0.470	0.003	21
CatBoost	0.601	0.530	0	17
KNN	0.556	0.447	0.014	20
MLP	0.573	0.491	0.0003	-

Table 5.7: Interception rating prediction performance of the models

As seen in Figure 5.16 an even smaller upward trend is observed combined with a lot more variability in the predictions versus the actual interception ratings scatter plot. The variability in predictions leads to the high RMSE and MAE values, while the extremely low R^2 indicate that the features almost don't have predictive power. Contrary to passes and dribbles, no real correlation is found between the test results and interception ratings. Unfortunately, no reliable conclusion can be drawn from these models. One possible reason for this could be that the passing/dribbling skills are evaluated using models that accurately estimated event difficulty, whereas interceptions could not be evaluated using such probabilistic models resulting in the usage of a more arbitrary metric. The metric can be improved by using more sophisticated models such as the one developed by Piersma (16). Another reason could be the size of the data sets, as interceptions is a more rare event, players often only have between 2-6 interceptions recorded. A sample size this small probably doesn't accurately capture a players interception ability.

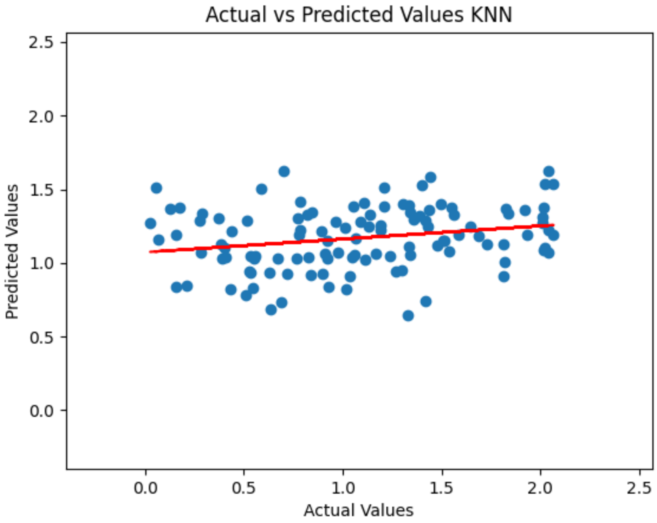

Figure 5.16: Actual vs predicted interception ratings KNN

6

Conclusion

In this study, the primary goal is to capture which skills and characteristics are crucial for youth players to possess to perform well in matches, assessed using objective and more sophisticated metrics than just success rate. By bridging the gaps in the current state of knowledge, the aim was to improve the identification/development of talent. The central research question of this study is 'To what extent can player characteristics and/or skills be used to predict the performance of football players?'

The difficulty of the events, expressed by xSP and xSD, are well estimated using a random forest and catboost respectively. Despite utilizing the same models for the similar problems, no model was superior for both and different models excelled. For passes, the random forest achieved a brier score of 0.171 with a calibration almost perfectly following the line x=y. The most important features were the length and angle of the pass combined with the applied pressure and position on the field. The data used was scaled based on age and level while no PCA or SMOTE oversampling was applied despite the class imbalance. Slight oversampling to a 0.25:0.75 ratio was optimal for xSD estimation, where catboost achieved a brier score of 0.109 and a slightly worse calibration. The lower brier score can be explained by the bin imbalance of the dribbles, as very few dribbles are predicted below a 50% success probability. The number of teammates behind the ball after the dribble has the most predictive power of dribble difficulty, indicating that not only position on the field is important but context matters as well. The highest speed during the dribble and applied pressure also played a significant role in accurately estimating dribble difficulty. Integrating test features for individual event success prediction decreased model performance, meaning that no information regarding the influence of characteristics on specific event types is obtained.

Referring back to the research question, the test results could not be used to explain the entire disparity in performance differences. This does not mean that nothing is achieved, for passing and dribbling correlations were observed, indicating that players with certain skills tend to perform slightly better. The best model for PPR prediction was XGBoost, achieving a R^2 of 0.175. Interestingly, the fifteen features used by the XGBoost were mental strength/attitude, devotion and cognitive ability. Sub-scores of the PUP10 test did not appear, thus a player's passing ability can not be assessed by a simple passing test, the essential elements lie in the mind. The KNN exhibited the best performance in predicting DPR with features that are significantly more important than other features. Contrary to passing, dribble capabilities mainly rely physical characteristics: biological age and ball control speed. Some players mature at a later age, it is too premature to dismiss their potential. Conversely, players that mature earlier could quickly reach their peak. As seen in Chapter 2, studies often overlook the biological age of youth players, yet the significance becomes apparent. The interception rating was calculated using a more subjective weighted average due to the lack of resources to develop a more sophisticated metric. Together with the significantly less interception occurrences, this can be part of the reason that no correlation could be found in the interception rating and the test results. Out of the six models used, four of them have exhibited the best performance at least once, stressing the importance of adopting a diverse selection of models with their respective (dis)advantages.

As discussed in Chapter 7, the subpar test data quality impacted the results. More extensive and complete data sets could potentially lead to more, improved insights.

7 Discussion

7.1 Limitations

7.1.1 Data quality

The limitations of this research are primarily in the quantity and quality of the data. First, there is not a player that has participated in all the tests conducted, which means that approximately 50% of the test data consists of NaN values. Even though these missing values are imputed, the imputation quality is low given the amount of missing data. Furthermore, each player performs the tests only once, introducing variability. A singular test result does not provide a comprehensive reflection as numerous circumstantial factors can influence the test score, causing inaccurate reflections of the player's skill set. When conducting the tests myself, some players required time to familiarize with certain exercises before performing it at their best. Due to the tight training schedules of the clubs, often there was not enough time for players to achieve their best scores. Another thing that raises concern about the test data quality is that the questionnaires about mental skills are subject to response bias and dishonest answers. It was noticed that players occasionally asked about the implications of the tests on the decision making about their future at the club, indicating a possible fear of giving honest answers to the questions of the questionnaires. Although the event data sets were of higher quality, the size is a point of consideration. Similarly to the test data, only one or two matches worth of data are available per player. To fully assess a player's abilities in matches, it is imperative to record data over more games, as not a single player consistently performs well every single game he plays, resulting in a skewed view.

As discussed, these concerns about the data quality could be a significant reason the test data does not have a higher predictive power. When Forward Football obtains more and data with a higher quality, the pipeline made in this study can be used again to gain new and improved insights.

7.1.2 Metrics

As mentioned in Chapter 2, there are opportunities for improvement of the weighted average used for the interception rating. Specifically, implementing a similar model proposed by Piersma (16) that takes interception difficulty and missed interception opportunities into account would improve results.

7.2 Suggestions for further research

For future research, it is recommended to address the limitations regarding the data and possibly try a wider range of models. Instead of implementing more models, ensemble methods could be explored, making the predictions a weighted average of the models already implemented. Although this study focuses on the present, the objective of identifying talent is to identify players that perform at a high level in the future. A skill that is not highlighted in this paper as it did not determine which players perform well right now could be crucial in later stages. It is not known whether players with certain characteristics and skills perform at a high level at a later stage. Future studies with substantial resources could research this by tracking groups of players over an extended period of time, continuously conducting tests and recording matches to observe changes throughout the time span and fully document development trajectories. Especially after the observation that the biological age is a significant factor in the performance of youth players, it could be valuable to explore the trajectory how players with varying biological ages eventually develop. Analyzing the trajectory of players over a longer period of time would provide more context to which players show true potential.

Overall, the field of talent identification/development still presents numerous opportunities for further exploration, not only in football but across all sports. With the ever-changing world of sports and advancements in technology, new possibilities to approach these challenges will present.

References

[1] D. Alaminos, I Esteban, and M. A. Fernández-Gámez. **Financial Performance Analysis in European Football Clubs**. *Multidisciplinary Digital Publishing Institute*, 2020. 1

[2] D. Miragaia, J. Ferreira, A. Carvalho, and V. Ratten. **‘Interactions between financial efficiency and sports performance Data for a sustainable entrepreneurial approach of European professional football clubs’**. *Emerald Insight*, 2019. 1

[3] V Ratten. **‘Developing a theory of sport-based entrepreneurship’**. *Multidisciplinary Digital Publishing Institute*, 2010. 1

[4] Sky Sports. **Kevin De Bruyne breaks down his most iconic PL assists**, 2023. 5

[5] G. Anzer and P. Bauer. **‘Expected passes: Determining the difficulty of a pass in football (soccer) using spatio-temporal data’**. *Data Mining and Knowledge Discovery.*, 2022. 5

[6] A. Fairchild, K. Pelechrinis, and M. Kokkodis. **‘Spatial analysis of shots in MLS: A model for expected goals and fractal dimensionality’**. *Journal of Sports Analytics .*, 2018. 5, 35

[7] P. M. Pardo. **‘Creating a Model for Expected Goals in Football using Qualitative Player Information’**. *Universitat Politècnica de Catalunya (UPC) - BarcelonaTech.*, 2020. 6

[8] M. Cavus and P. Biecek. **‘Explainable expected goal models for performance analysis in football analytics’**. 2022. 6

[9] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. **‘A comparative analysis of gradient boosting algorithms’**. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, 2021. 6

[10] E. Al Daoud. **'Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset '**. *World Academy of Science, Engineering and Technology.*, 2019. 6

[11] R. Szczepanek. **'Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost'**. *Hydrology.*, 2022. 6

[12] I. Dick, M. Tavakol, and U. Brefeld. **'Rating Player Actions in Soccer'**. *Front. Sports Act. Living.*, 2021. 7, 21

[13] T. Decroos, L. Bransen, J. van Haaren, and J. Davis. **'Actions Speak Louder than Goals: Valuing Player Actions in Soccer '**. *Applied Data Science Track.*, 2019. 7

[14] C. Trainor. **'Defensive Metrics: Measuring the Intensity of a High Press '**. *Statsbomb.*, 2014. 7

[15] C. Merhej, R.J. Beal, T Matthews, and S. Ramchurn. **'What Happened Next? Using Deep Learning to Value Defensive Actions in Football Event-Data'**. In *27th ACM SIGKDD Conference on Knowledge Discovery Data Mining.*, 2021. 8

[16] Piersma, J.P.T. *Valuing Defensive Performances of Football Players.* Master's thesis, Erasmus University Rotterdam, 2020. 8, 44, 50

[17] T. McGuckian, A. Beaven, J. Mayer, and D. Chalkley. **The association between visual exploration and passing performance in high-level U13 and U23 football players**. *Science and Medicine in Football*, 2020. 8

[18] M. Bila and A.R. Hillman. **The Relationship Between Anxiety, Confidence and Short-Passing Performance in Collegiate Soccer Players**. *Research Directs*, 2021. 8

[19] P. Lipinska and A. Szwarc. **'Laboratory tests and game performance of young soccer players'**. *Trends in Sport Sciences*, 2016. 8

[20] M.R. Abdullah, R.M. Musa, A.B.H.M.B. Maliki, N.A. Kosni, and P.K. Suppiah. **Role of psychological factors on the performance of elite soccer players**. *Journal of Physical Education and Sport*, 2016. 8

[21] A. L. Kelly, C. A. Williams, R. Cook, S. L. J. Sáiz, and Wilson M. R. **'A Multidisciplinary Investigation into the Talent Development Processes at an English Football Academy: A Machine Learning Approach'**. *Multidisciplinary Digital Publishing Institute*, 2022. 8

[22] H. Sarmento, M.T. Anguera, A. Pereira, and D. Araújo. **'Talent Identification and Development in Male Football: A Systematic Review'**. *Sports Medicine.*, 2018. 9

[23] D. Fortin-Guichard, I. Huberts, J. Sanders, R. van Elk, D.L. Mann, and G.J.P. Savelsbergh. **'Talent Identification and Development in Male Football: A Systematic Review'**. *Sports Performance.*, 2022. 9

[24] T. de Joode. **'The Football Skills Track as an objective measurement tool to identify talented youth (9-12 years old) regarding (ball) movement skills'**. *Department of Human Movement Sciences, Vrije Universiteit Amsterdam.*, 2017. 12

[25] O. BenOunis, A. BenAbderrahman, K. Chamari, A. Ajmol, M. BenBrahim, A. Hammouda, M. Hammami, and H. Zouhal. **Association of Short-Passing Ability with Athletic Performances in Youth Soccer Players**. *Asian Journal of Sports Medicine*, 2013. 12

[26] A. Ali, C. Williams, M. Hulse, and A. Strudwick. **'Reliability and validity of two tests of soccer skill'**. *Journal of Sports Sciences*, 2007. 13

[27] J. Côté and B. Baker, J. abd Abernethy. **'From play to practice: A developmental framework for the acquisition of expertise in team sport.'**. *Champaign, IL: Human Kinetics.*, 2003. 14, 42

[28] L. Peng and L. Lei. **'A Review of Missing Data Treatment Methods'**. *Department of Information Systems, Shanghai University of Finance and Economics.*, 2005. 16

[29] J. Huanga, J.W. Keunga, F. Sarro, Y. Li, Y.T. Yua, and Hongyi Sund Chana, W.K. **'Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study'**. *The Journal of Systems and Software*, 2017. 17

[30] A. Lall. **'Data streaming algorithms for the Kolmogorov-Smirnov test.'**. *E International Conference on Big Data (Big Data).*, 2015. 17

[31] H. de Silva and A.S. Perera. **'Evolutionary k-Nearest Neighbor Imputation Algorithm for Gene Expression Data'**. *International Journal on Advances in ICT for Emerging Regions*, 2017. 17

[32] R. Salmerón, C. B. García, and J. García. **'Variance Inflation Factor and Condition Number in multiple linear regression'**. *Journal of Statistical Computation and Simulation*, 2017. 18

[33] S.Q. Lafi and J.B. Kaneene. **'An explanation of the use of principalcomponents analysis to detect and correct for multicollinearity '**. *Preventive Veterinary Medicine*, 1992. 18

[34] M.O. Akinwande, H.G. Dikko, and A. Samson. **'Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis'**. *Open Journal of Statistics*, 2015. 18

[35] S.B. Franklin, D.J. Gibson, P.A. Robertson, J.T. Pohlmann, and J.S. Fralish. **'Parallel Analysis: a method for determining significant principal components'**. *Journal of Vegetation Science*, 2015. 18

[36] S. Vijendra and P. Shivani. **'Robust Outlier Detection Technique in Data Mining: A Univariate Approach '**. *Mody Institute of Technology and Science*, 2011. 19

[37] J. Brandt and E. Lanzén. **'A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification'**. *Department of Statistics Uppsala University*, 2020. 21

[38] X. Wang, J. Gong, Y. Song, and J. Hu. **'Adaptively weighted three-way decision oversampling: A cluster imbalanced-ratio based approach'**. *Applied Intelligence*, 2023. 21

[39] X. Chen and H. Ishwaran. **'Random forests for genomic data analysis'**. *Division of Biostatistics, Department of Epidemiology and Public Health, University of Miami.*, 2012. 23

[40] L. Breimann. **'Random Forests'**. *Statistics Department, University of California, Berkeley*, 2001. 23

[41] C. Zhang, L. Cao, and A. Romagnoli. **'On the feature engineering of building energy data mining'**. *Sustainable Cities and Society.*, 2018. 24

[42] J. H. Friedman. **'Greedy function approximation: a gradient boosting machine'**. *The Annals of Statistics.*, 2001. 25

[43] A Natekin and A. Knoll. **'Gradient boosting machines, a tutorial'**. *Frontiers in Neurorobotics.*, 2013. 25

[44] J. Ramzai. **'Simple guide for ensemble learning methods'**. *Towards Data Science.*, 2019. 25

[45] T. Chen and C. Guestrin. **'XGBoost: A Scalable Tree Boosting System'**. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, 2016. 27

[46] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.Y. Liu. **'LightGBM: A Highly Efficient Gradient Boosting Decision Tree'**. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, 2016. 27

[47] E. Fix and J. Hodges. **'Explainable expected goal models for performance analysis in football analytics'**. 1951. 28

[48] Z. Yao and W. Ruzzo. **'A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data'**. *BMC Bioinformatics.*, 2006. 28

[49] C. Hu, G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka. **'Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithiumion battery'**. *Appl Energy.*, 2014. 28

[50] M.M. Kumbure and P. Luukka. **'A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance'**. *Granular Computing.*, 2021. 28

[51] A. Krogh. **'What are artificial neural networks?'**. *Computational Biology.*, 2008. 29

[52] K. Hornik, M Stinchcombe, and H. White. **'Multilayer Feedforward Networks are Universal Approximators '**. *Neural Networks.*, 1998. 29

[53] M.W. Gardner and S.R. Dorling. **'Artificial neural networks (the multilayer perceptron) —a review of applications in the atmospheric sciences'**. *Atmospheric Environment.*, 1998. 29

[54] V.H.A. Ribeiro, G. Steinhaus, E.B. Severo, J.R. Ferreira Junior, L. Jos, L. Barbosa, M. Cossetin, M. Vinicius, and M. Figueredo. **'A System for Enhancing Human-level Performance in COVID-19 Antibody Detection.'**. In *Simpósio Brasileiro de Computação Aplicada à Saúde*, 2021. 29

[55] J.D.R. Olvera, I. Gómez-Vargas, and J.A. Vázquez. **'Observational Cosmology with Artificial Neural Networks'**. *Universe* ., 2022. 30

[56] Rumelhart D.E., Hinton G.E., and Williams R.J. **'Learning internal representations by error propagation'**. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition.*, 1986. 30

[57] M. Gori and A. Tesi. **'On the Problem of Local Minima in Backpropagation'**. *Transactions on pattern analysis and machine intelligence.*, 1992. 31

[58] K.A. Carpenter, D.S. Cohen, J. Jarrell, and X. Huang. **'Deep learning and virtual drug screening'**. *Future Medicinal Chemistry.*, 1986. 31

[59] K. Rufibach. **'Use of Brier score to assess binary predictions'**. *Journal of Clinical Epidemiology*, 2010. 32

[60] D.J. Spiegelhalter. **'Probabilistic prediction in patient management and clinical trials'**. *Statistics in medicine.*, 1986. 32

[61] F. Barili, D. Pacini, A. Capo, E. Ardemagni, G. Pellicciari, M Zanobini, C. Grossi, K.M. Shahin, F. Alamanni, R Di Bartolomeo, and A. Parolari. **'Reliability of New Scores in Predicting Perioperative Mortality After Isolated Aortic Valve Surgery: A Comparison With The Society of Thoracic Surgeons Score and Logistic EuroSCORE'**. *The Annals of Thoracic Surgery.*, 2013. 33

[62] Y. Huang. *'Preservation of Patient Level Privacy: Federated Classification and Calibration Models'*. PhD thesis. 33

[63] O. Lindhiem, I.T. Petersen, and E.A. Mentch, L.k.and Youngstrom. **'The Importance of Calibration in Clinical Psychology'**. *Sage.*, 2020. 33

[64] FBRef. **2022-2023 Premier League Miscellaneous Stats**, 2023. 40

Appendix A

Feature	Meaning	Range
APHV	Age of peak height velocity	12.53-15.81
MO	Maturity Offset, years until peak height velocity	-2.43-2.95
Ballcontrol speed	Time 8 figure + backwards dribble part of the FST 3.1	3.71-10.82
Dribble speed	Time of dribble + slalom part of the FST 3.1	16.77-34.95
FST Final time	total time completing the FST 3.1 + penalty	24.04-40.17
PUP10 Time	Time completing passing course including penalty for wrong passes	21.29-50
Serie	Which foot which the player should use for specific passes in PUP10	categorical
Decision Making	Processing speed of visual information	25-70
Eye Movement	Hand-eye/foot-eye coordination measured by the RightEye vision system	39-70
Focus	Test that checks whether the scores of the previous two features change when adding distractions	62-100
Sports Functional Score	Ability in spotting contrasts in figures	57-84
Sports Vision	Football specific Eye Vision score	51-74
Best trial	Best score out of 3 VFMT tests, see Equation 3.1	5.51-21.58
Recognizing Body language	Correct ball trajectory section answers	12-27
Recognizing Ball trajectory	Correct ball trajectory answers	1-12
Positioning Speed	The time a player chooses the trajectory of the ball	0-16
GiT Score	Final score calculated by adding up the three features above	14-47
IJBT ballcontrol Score	Scaled average of toe taps and hat dance count	0.26-0.99

IJBT feint Score	Scaled average of remaining IJBT test counts	0.09-0.85
IJBT Final Score	Scaled average of all counts	0.16-0.82
IJBT ballcontrol Fluency	Scaled average of toe taps and hat dance fluency	0.21-0.99
IJBT feint Fluency	Scaled average of remaining IJBT test fluencies	0.29-0.89
IJBT Fluency Score	Scaled average of all fluencies	0.34-0.93
LSPT3 Final time (total seconds)	End time LSPT + penalty seconds	28.16-98
Score Spatial	Correct figure identifications in a group of four figures	0-17
Score Memory	Quick Correct figure identifications in a group of four figures	1-16
Score Seq Memory	Correct sequence identifications in a group of four figures	0-16
Vis Clo	Correct identifications of objects when only a part is shown	0-16
Vis Fig Ground	Correct figure identifications in a group of four figures	2-16
Final TVPS3 Score	Average of the five features above	3.33-16
M-Compact total score		311-508
MQ-score	Scaled value of motor ability tests	84-127
Mental Entity	Questions about how much talent determines ceiling	1-5
Mental Incremental	Questions about how much training determines ceiling	1-5
Confidence	Questions about the confidence of the player	1-5
Control	Questions about how the player deals with his emotions	1-5
Challenge handling	Questions about how the players deal with possible challenges	1-5
Commitment	Questions about persistence to achieve goals	1-5
Learning Orientation	Questions about eagerness of learning	1.17-5
Performance Orientation	Questions about eagerness of performing	1.33-5
Focus playing vs training	Questions on how much the player is focused on playing or performing	1.34-4.5
Number of sports played	Total number of sports the player has played	1-4
Total hours guided training	Cumulative hours of football training	260-12345

Total hours free play	Total hours the player has played sports excluding training	104-9776
Percentage football Age 4-12	Percentage of sports time dedicated to playing football between age 4 and 12	0.3-1
Percentage football Age 12-15	Percentage of sports time dedicated to playing football between age 12 and 15	0.33-1
Percentage football Age 16+	Percentage of sports time dedicated to playing football between age 16+	0.38-1

Table A.1: Explanation of the used test features and their respective ranges of recorded values

Feature	**Meaning**	**Range**
Time block	At which part of the match the pass happened, each match is divided in six blocks	1-6
Zone	Whether the pass is in defence/midfield/attack	-
Area	The are on the pitch, the pitch is divided in 18 different areas	1-18
Angle	The angle of the pass, expressed in an absolute value of π	0-3.14
Pass type	Whether the pass was lateral/backwards/forward	-
Pass length	The length of the pass in meters	1-76.84
adj_X	Adjusted X coordinate of the player giving the pass based on playing direction	0-106
adj_Y	Adjusted X coordinate of the player giving the pass based on playing direction	0-68
Pressure direction	The direction from where the player is pressured, back or front	-
Distance 2 opp	Distance to the closest opponent in meters	0.63-47.50
Mean distance	Mean distance towards all opponents on the pitch	4.88-51.63
Distance front	Closest opponent in front of the player	0.2-100
Distance back	Closest opponent behind the player	0.2-108.6
Pressure level	Categorical variable rating the pressure as no/limited/full pressure	-
Pressure with direction	The lowest value of distance back/front with distance back adjusted to a negative value	-29.12-40.40
Pressure 10	Pressure level expressed on a scale between 0-10	0-10

Table A.2: Explanation of the used pass features and their respective ranges of recorded values

Feature	Meaning	Range
ivt_X_begin	Adjusted X coordinate of place where the dribble starts based on playing direction	0-105
ivt_Y_begin	Adjusted Y coordinate of place where the dribble starts based on playing direction	0-69
ivt_X_end	Adjusted X coordinate of place where the dribble ends based on playing direction	0-108
ivt_Y_end	Adjusted Y coordinate of place where the dribble ends based on playing direction	0-71
Dribble Time	The duration of the dribble in seconds	1.2-57.8
Dribble Distance	The distance the player covered with the ball	0.50-71.44
Highest Dribble Speed	Highest speed recorded during the dribble in m/s	0-10
NearbyTeammates		311-508
Opponents in 1m radius	Average opponents within 1m during the dribble	311-508
AverageOpponents	Average opponents within 5m during the dribble	311-508
MaximumOpponents	Maximum opponents within 5m during the dribble	311-508
Teammates before ball begin	Teammates behind the ball before the dribble	0-10
Teammates before ball end	Teammates behind the ball after the dribble	0-10
Opponents before ball end	Opponents behind the ball before the dribble	0-11
Opponents before ball begin	Opponents behind the ball after the dribble	0-11
Total playing time	Total seconds the player has been on the pitch	1.2-5476
xT begin	Expected threat of the coordinates where the dribble started	0-0.184
xT end	Expected threat of the coordinates where the dribble ended	0-0.184
ATC dribble	Attacking Threat Contribution; difference between xT begin and xT end	-0.144-0.138
Zone begin	The pitch is divided in four zones (horizontal lines), this denotes the starting zone	-
Zone end	The pitch is divided in four zones (horizontal lines), this denotes the ending zone	-
IsDribbleForward	A boolean on whether the dribble is forward	-

Table A.3: Explanation of the used dribble features and their respective ranges of recorded values

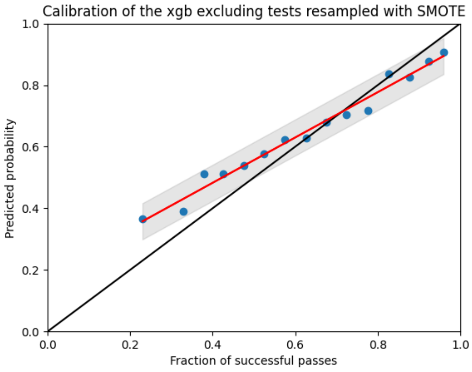

Figure A.1: Calibration plot XGB with pass features and SMOTE oversampling

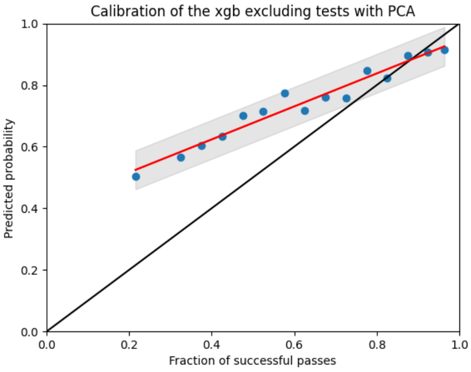

Figure A.2: Calibration plot XGB with pass features and PCA