Bachelor's Thesis

Analysis and Prediction of Successful Football Player Transfers

Department of Statistics
Ludwig-Maximilians-Universität München

Justin Lampman

Munich, March 24th, 2025

Submitted in partial fulfillment of the requirements for the degree of B. Sc.
Supervised by Prof. Dr. Thomas Nagler

Abstract

This thesis aims to predict soccer transfers in the men's European top five leagues, using data from the 2017 to 2024 season. Using principal component analysis (PCA) to categorize the playing styles among teams, Random Forest (RF) are used to predict a gain in market value. The prediction model was applied to both a binary classification task and a regression task, with the main focus on binary classification. A successful transfer is defined not just by an increase in market value, but also adjusted for a players tenure at a club, his age and the expected depreciation of value over time. The analysis identified that team-based statistics have a greater influence on a successful transfer than player-based statistics, while the players position was comparatively less important. Differing team playing styles between old and new clubs play a crucial role in determining successful transfers with their impact varying across different positions.

Contents

1 Introduction **1**

2 Previous Work **2**

3 Data **4**

4 Successful Transfer Definition **5**

5 Theory **8**

5.1 CART Model 8

5.2 Random Forest 9

5.2.1 Variable Importance 11

5.2.2 Variable Interaction 11

5.3 Principle Component Analysis 12

5.3.1 Variance Maximization 13

5.4 Partial dependence plots 14

6 Results **16**

6.1 Principal Component Analysis 16

6.2 Random Forest 18

6.2.1 All Position Model 19

6.2.2 Forward Model 25

6.2.3 Midfield Model 28

6.2.4 Defender Model 32

6.2.5 Similarities between all Models 36

7 Conclusion **40**

A Appendix **VIII**

A.1 Variables imported from FBref and Transfermarkt VIII

A.2 Additional PDP for all models XXII

A.3 Variables used for PCA XXX

B Electronic appendix XXXII

1 Introduction

The world we live in today places great importance on sports. Be it for exercising and clearing the head, practicing sports for health reasons, watching sports for enjoyment, or being united as one nation; we find sports in many aspects of our life. We see the growing importance for sports in our life, when we look at economic output of the global sports industry generating over 400 Billion Dollars a year in 2022 alone (Gough, 2024).

Since sports teams, leagues, and enterprises are valued so highly in our society, it becomes a lucrative business opportunity to generate wealth. With money at stake, making the right decision is important for the well-being of the the soccer club or institution. With the biggest sport in the world being soccer and a large part of every soccer club being signing players, making the right or wrong decision of signing a player, can not only win you the season, but could also throw the club into turmoil for years to come (Dvorak et al., 2000).

To address the challenge of making the right decision, the coaching staff implements a wide range of strategies to predict if a player will fit into the squad, or what player one should look for. One such approach utilizes the tree-based methods with Random Forests (Bartosz et al., 2021).

The contribution this thesis makes is a possible implementation for machine learning techniques to predict a successful transfer of a soccer player for professional soccer clubs and distinguish the most valuable features for a player position. In addition, a new framework for team-based playing styles is implemented into the model to account for differing playing styles across teams. For each club, player and game, a vast amount of data is retrieved and collected. While some variables are important for deciding if a transfer will be successful, others are not. The aim of this work is to distinguish between the variables predictive for the success of a player transfer. The thesis is structured as follows: Section 2 discusses previous work and related fields for Machine Learning applications. In section 3 we will discuss the data, followed by section 4 where a successful transfer is defined. Section 5 dives into the theory used in this work. Section 6 examines the results and ending the thesis in section 7 with a sumary and outlook.

2 Previous Work

Sports analytics has become a widespread phenomenon not just among professional sports teams, coaches or odds makers, also recreational sports use the growing technology to analyze, predict, or better ones ability to play the game. Nowadays a large number of structured, as well as unstructured data is produced which opens up a new world to view and understand the activity and sport. This data is used for a wide range of applications, one major use of which is the prediction of future events. With growing technological ability, machine learning approaches become increasingly popular in the world of sports, which aim to assist managers, coaches, and players in decision making, predicting results, player injuries, performance or scouting.

One of the earliest adoptions and uses of sports analytics was the prediction of outcomes of sports games or gaining insights into the underlying dynamic of the game (Puerzer, 2002). Early approaches used to predict outcomes of games relied on mathematical and statistical approaches which were often verified by an expert of the field (Bunker and Thabtah, 2019). Newer approaches have used prediction models based on machine learning, which are also used in other fields, including medical diagnoses (Bhavsar et al., 2021), finance (Ludkovski, 2023), or biology (Rai et al., 2024). Sports analytics has progressed with not only outcomes being measured and predicted. A growing importance in sports is being put on other variables, such as injury prevention, which is increasingly modeled with Machine Learning algorithms as well (Wei et al., 2023). For this topic, a wide range of approaches are used to predict results and gain insight with models ranging from random forests, decision trees, neural networks, to logistic regression.

Machine learning is not limited to a single use or sports, but has been implemented in a wide range of topics. Machine learning has been used in American football to predict defensive positioning (Newman et al., 2023), in handball for prediction of certain movements (Lentz-Nielsen et al., 2023), in basketball to evaluate training output (Pengyu and Wanna, 2021), in ice hockey for evaluating actions on ice (Schulte et al., 2017), also in soccer for the prediction of injury risk (Shen et al., 2023) and many other sports and applications. Sports betting analysis has also been done with the increasing popularity in this field, with over eleven billion dollars being spent on sports betting in the U.S.

alone in the year 2023 (Knoll and Stübinger, 2020). With almost every aspect of sports able to be modeled with machine learning, sports federations have also made efforts to automate refereeing with such methods using neural networks and labeled video data to ensure unbiased and efficient refereeing (Ma and Kabala, 2024).

This work focuses on an approach for scouting, transfer aid and predictions. There are many tools to gather data for scouting purposes, however they seldom provide transfer suggestions, which is still a very human and subjective field. Efforts were made for the NHL draft, predicting draft outcomes based on previous college performances (Luo, 2024). Similar approaches have been made in soccer, with efforts to predict football player ratings (Bhatnagar et al., 2024).

The biggest step towards the goal of transfer prediction with machine learning in soccer has comes from Bartosz et al. (2021) that this work builds on. Here, player data gets grouped into four parameters: physical, technical, psychological and age. These are then assessed based on three different definitions of a successful transfer.

While machine learning plays an ever growing role for models in any field, for soccer teams, attempts have been made outside machine learning to understand the sport in a more profound way. One such attempt was made by Kong et al. (2022) using principal component analysis (PCA) to analyze a team's playing style. Further work has been done in this field, optimizing this approach to identify team playing styles in soccer (Plakias et al., 2023)

3 Data

The player match data is gathered from Stathead.com (2025), which publishes data related to a variety of sports; mainly football, baseball, basketball and hockey. In this work the data is focused on the top five leagues in European soccer, which are the first division leagues in Spain, Germany, England, France and Italy, as well as the seasons 2017/18 through 2023/24. The transfers considered are between the top five leagues, which means that transfers from and to outside of these leagues were not considered. In addition, this work focuses only on outfield players and excludes goalkeepers. The position of the goalkeeper in soccer is very unique, with every team only being able to field one at a time and having a special role and therefore also unique metrics. Goalkeepers therefore would have been assessed based on different metrics while also having only 176 transfers in the time span. For this, reason goalkeepers were excluded from the analysis. For outfield players, metrics were categorized into nine different categories: Standard Stats, Shooting, Passing, Pass Types, Goal and Shot Creation, Defensive Actions, Possession, Playing Time and Miscellaneous Stats. Combined these metrics result in over 170 statistics. For the market values and the positions of the players, data was gathered from Transfermarkt.com (2025) and combined with the data from Stathead.com (2025) to ensure that an accurate market value is assigned to the player. Market values are changed throughout the year. To prevent this from affecting the analysis the market values used are all from the end of the season. The final dataset includes 2679 transfers.

4 Successful Transfer Definition

To define whether a transfer was successful or unsuccessful, a metric is used which continuously calculates the gain or loss in market value over the period employed at a club. For this purpose the following formula is postulated:

$$TransferValue_i = -x_1\beta_1 + \sum_{i=2}^{n_{years}} \frac{x_i\beta_i}{i \sum_{k=2}^{n_{years}} \frac{1}{k}} \tag{1}$$

This formula uses the player's age to calculate the transfer value, with β_1 being the average loss or gain in market value for a player at that age, while β_{1+i} is that coefficient after i years at a team. The market value of a player is denoted as x_i, with x_1 being the market value of a player at the time of the transfer. Here, n_{years} denotes the years a player is at a team after a transfer. For a successful transfer to occur, we require that the starting market value adjusted for age $x_1 \cdot \beta_1$ has to be overcome by the adjusted market value gained or lost during the tenure at a team. Transfer cost and the value spent by the club were not accounted for, since the actual transfer cost is based on multiple non-quantifiable metrics such as demand or the need to get a player in time. The deficit is overcome if a player has a positive return in market value adjusted for age over the tenure at the club. The years are weighted by the duration at the team, to ensure that the impact shortly after a transfer is valued more than the impact after ten years at the team. Teams are often impatient, wanting to see results sooner than later, which is why it is important for teams to have an impact sooner than later. Each year after a transfer is weighted less, so that each year k after a transfer is weighted in the formula as follows:

$$\frac{1}{i \sum_{k=2}^{n_{years}} \frac{1}{k}}. \tag{2}$$

This decay is based on the harmonic series, adjusted so that all weights sum to one. In addition to this, the median transfer value was calculated for each age, to account for the average loss of market value after the mid-20s. As can be seen in Figure 1, the median transfer values change for different ages.

For a transfer made at the age of 19, the median value of a transfer is 2.8 € Million, while

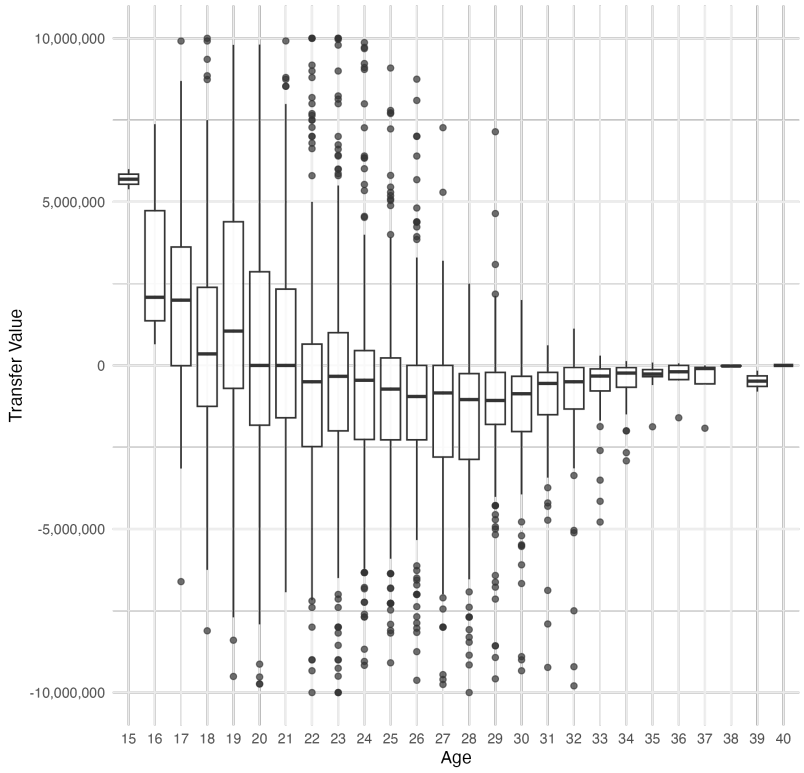

Figure 1: Median value per age, cut off at values -10 Million and 10 Million, for better visualization, representing about 90% of of transfers

the median for a transfer of a 28 year old is at about -1.3 € Million. With this in mind, a successful transfer is defined in two ways: for the regression model, the normal transfer value as defined in formula 1 is used. For the binary classification task, a successful transfer is denoted as 1 and defined as:

$$ST_i = \begin{cases} 1 & if \quad TransferValue_i > Median(TransferValue \mid Age_i), \\ 0 & else \end{cases} \tag{3}$$

Therefore, a successful transfer (ST) is defined as the transfer value that exceeds the median transfer value for that age. This definition is more strongly adjusted to the disproportionate effect age has on market value and is therefore more robust against the age of a player compared to definition 1. Through the binary definition of a transfer, half of transfers were classified as successful while the other half below the median was classified as a unsuccessful transfer. Since both the age and the market value of a player are used to define a successful transfer, these variables were excluded from the variables

used in the models to determine a successful transfer.

5 Theory

The amount of data generated from soccer matches is substantial. With over 170 variables per player alone, it is difficult to distinguish which variables or metrics are significant for a players transfer and which have no importance at all. With a regression or a binary classification task, lasso or ridge regression is an option to select features. Due to the robustness of random forests, this method was used to determine feature importance (Kursa, 2014).

5.1 CART Model

The foundation of Random Forests is the Classification and Regression Tree (CART) method. This method was formally introduced by Gordon et al. (1984) as a decision tree that splits data into homogeneous subsets using a predefined splitting criterion. A decision tree is a mapping function for both regression and classification tasks, which assigns each input vector $\mathbf{X}$ a predicted output $\hat{\mathbf{y}}$ while the learning set is defined as $\mathcal{L} = (\mathbf{X}, \mathbf{y})$ (Louppe, 2014). A specific split that a tree uses to divide a node into its descendants will be selected while considering every other splitting predictor variable, based on predefined splitting criterion (Cutler et al., 2012). This differs for regression and classification tasks. Regression tasks focus on minimizing the prediction error (usually with MSE), resulting in a piecewise constant model that divides the predictor space into regions where response variables are similar. Classification tasks, on the other hand, assign each region a class based on majority rule, while each split is determined by a predictor variable leading to two branches that continue until terminal nodes are reached. This split is usually done by Gini Index or impurity (Strobl et al., 2007).

Classification tasks, on the other hand, assign each region a class based on majority rule. Each split in the tree is determined by a predictor variable leading to two diverging branches. This continues until terminal nodes are reached, which happens after when splitting criterion is met, or further splits do not improve the model. This split is usually done by Gini Index or impurity (Strobl et al., 2007).

The differences in models for classification and regression are shown in Figure 2 and

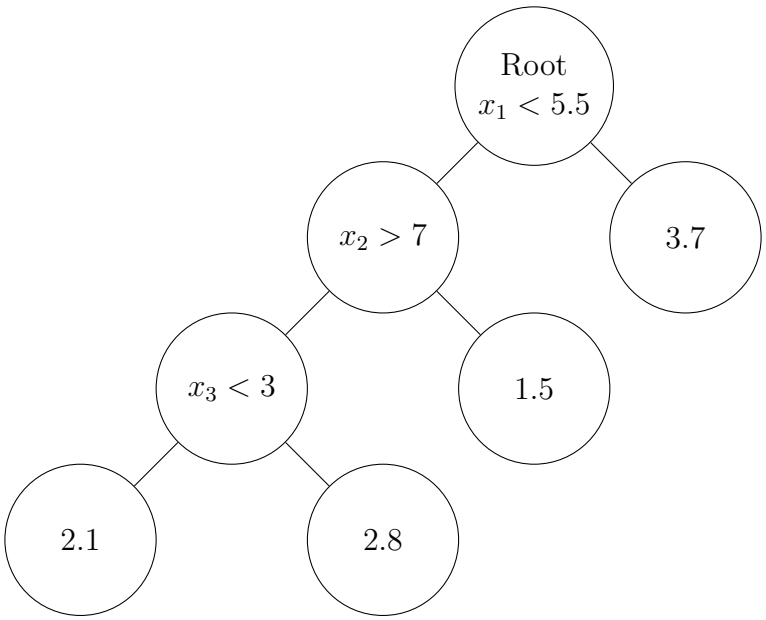

Figure 2: A simple Classification and Regression Tree (CART) for regression.

Figure 3. CART has the advantage that it is robust against extreme outliers, as well as its support to handle categorical as well as continuous features (Hastie et al., 2009).

5.2 Random Forest

Random forests are a concept developed by Breiman (2001), building on his previous work on CART. Random forest is an ensemble learning method that extends bagging while incorporating randomness of features in the tree building process, leading to improved and generalized performance compared to single decision trees (CART). This method can also be used for regression as well as classification problems, while preserving the same task definition as for CART. For both regression and classification tasks, this method provides robust handling of missing values as well as high-dimensional data (Cutler et al., 2012). As the task is the same as with CART, classification tasks predict using majority a vote

$$f(x) = \arg\max_{y \in Y} \sum_{j=1}^{J} I\left(y = h_j(x)\right) \tag{4}$$

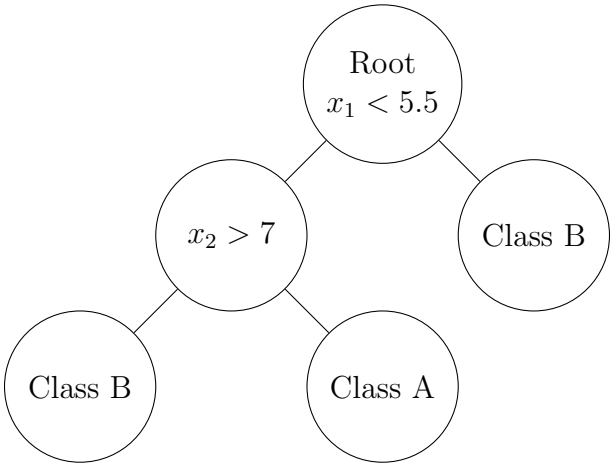

Figure 3: A simple Classification and Regression Tree (CART) for classification.

while for regression tasks, the prediction for a new observation is averaged

$$f(x) = \frac{1}{J} \sum_{j=1}^{J} h_j(x) \tag{5}$$

(Cutler et al., 2012). Here J stands for the number of trees in the forest, while j is the index for the individual tree in the ensemble and $h(x)$ represents the prediction of decision tree j. A random forest builds upon the structure of a CART, fitting multiple trees while using bootstrapped samples of the data and evaluating each node on a random subset of predictors to evaluate the best splitting criterion. The bootstrapped procedure generates a random sample of the data $D = \{(x_i, y_i)\}_{i=1,..n}$ and creates a new training set for D_b for each tree T_b that is fitted, while $b \in \{1, \ldots, P\}$ and P is the number of trees in the forest. This results in some observations being duplicate in D_b while others might be left out. Furthermore, each tree selects a random number of `mtry` variables to determine the split at each node in the tree. The process repeats until the minimum node size is reached, which results in the ensemble of trees called random forest. For this implementation the R package ranger by Wright and Ziegler (2017) was used for the training of the forests while caret by Kuhn and Max (2008), ModelMetrics by Hunt (2020) and pROC byRobin et al. (2011) were used for evaluating the models.

5.2.1 Variable Importance

The importance of different variables or features in a random forest can be quantified through feature importance, developed by Breiman (2001), which differs in a regression or classification setting. For classification problems, the feature importance is calculated with the unnormalized average importance of a feature x:

$$Importance(x) = \frac{1}{n_T} \sum_{i=1}^{n_T} \sum_{j \in T_i splitvar(j)=x} p_{T_i}(j) \Delta i_{T_i}(j). \tag{6}$$

Here, n_T is the number of trees and T_i depicts tree i. Whereas $p_{T_i}(j)$ is the fraction of samples reaching node j, while $\Delta i_{T_i}(j)$ is the change in impurity at node j in tree T_i.
One fundamental distinction between how the feature importance is computed in a classification and regression setting is that in classification, the Gini Importance is used as measure, whereas in regression, the focus is on reducing variance in the target variable. The variable importance for a regression task is therefore calculated using the variance reduction defined as:

$$I_V(N) = \tag{7}$$

$$\frac{1}{|S|^2} \sum_{i \in S} \sum_{j \in S} \frac{1}{2}(y_i - y_j)^2 - \left(\frac{|S_t|^2}{|S|^2} \frac{1}{|S_t|^2} \sum_{i \in S_t} \sum_{j \in S_t} \frac{1}{2}(y_i - y_j)^2 + \frac{|S_f|^2}{|S|^2} \frac{1}{|S_f|^2} \sum_{i \in S_f} \sum_{j \in S_f} \frac{1}{2}(y_i - y_j)^2 \right)$$

This formula quantifies the effectiveness of a split at node N and measures how much it reduces the variance for the target variable $I_V(N)$. S,S_t and S_f are a set of pre-split (S) and post-split (S_t, S_f)sample indices. The variable importance here measures how the variance is reduced (Breiman, 2001). By summing the variance reductions across all nodes that a feature has split, a global importance score is obtained that reflects the total contribution a feature has on reducing the prediction error. This allows to compare the influence of different variables on the model and assess which variable drives the predictive performance the most.

5.2.2 Variable Interaction

Variable interaction is a measure that assesses the strength of an interaction between pairs of variables in a predictive model. It compares the combined effect of two variables

to the sum of their individual effects and how these variables contribute to the model's prediction. Variable interaction can be quantified by Friedman's H-statistic, which is derived from Partial Dependence Plots which is explained in detail in section 5.4 (Friedman and Popescu, 2008). The formula is given as:

$$H_{jk}^2 = \frac{\sum_{i=1}^{n} \left[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \right]^2}{\sum_{i=1}^{n} PD_{jk}^2(x_j^{(i)}, x_k^{(i)})} \tag{8}$$

where $PD_{jk}(x_j^{(i)}, x_k^{(i)})$ represents the joint partial dependence and $PD_j(x_j^{(i)})$ as well as $PD_k(x_k^{(i)})$ represents the single partial dependence of the variables x_j and x_k. For this, the package vivid from Inglis et al. (2022) was used.

5.3 Principle Component Analysis

Principal component analysis (PCA) is a statistical method to reduce dimensionality of data, while trying to preserve as much variance as possible. This method can be traced back to Pearson (1901), and was further developed by Hotelling (1933). PCA is an unsupervised learning method and is often used for noise reduction or exploratory data analysis. The main goal of PCA is to reduce a dataset to a smaller set of variables called principal components (PC) that capture most of the variability of the original data. This is particularly useful for high dimensional data, reducing it to a more interpretable form. We now consider we have a matrix $\mathbf{X}$ with n observations and p numerical variables, whereas each column p corresponds to one variable and each row n corresponds to an observation. We also assume that $p < n - 1$ to ensure the variable estimates are well defined (Jung and Marron, 2009). Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be the observation vectors, while $\bar{\mathbf{x}} \in \mathbb{R}^p$ is the mean vector of these variables. In the standard approach, PCA is performed on the centered matrix $\mathbf{X}$ by subtracting the mean of each variable p so that $\mathbf{X}^* = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^\top$, where $\mathbf{1}$ is a vector of length n filled with ones. Hereby the centering ensures that each variables mean is zero, to result in the covariance being captured correctly by the products (Kim and You, 2023). Often, the variables p are measured on different scales or have incompatible units, in which case one may additionally standardize each column to unit variance (1);

PCA can then be preformed on the correlation matrix rather then the covariance matrix:

$$\mathbf{Z} = \mathbf{D}^{-1}\left(\mathbf{X} - \mathbf{1}\overline{\mathbf{x}}^{\top}\right) \tag{9}$$

where $\mathbf{D} = diag\left(\sigma_1, \sigma_2, \ldots, \sigma_p\right) \in R^{p\times p}$ is a diagonal matrix which contains the standard deviations σ_j of each variable, where σ_j is defined as:

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_{ij} - \overline{\mathbf{x}}_j\right)^2}, \quad j = 1, \ldots, p. \tag{10}$$

Contrary, the sample covariance matrix of p variables is defined as:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^{\top}, \tag{11}$$

where $\mathbf{S}$ is a $p \times p$ symmetric matrix of the covariances of each variable (Kim and You, 2023). Here the j and k elements of $\mathbf{S}$ depict the covariance between variable j and variable k, and the diagonal elements $\mathbf{S}_{1,1}...\mathbf{S}_{p,p}$ are the variances of each variable.

5.3.1 Variance Maximization

The principal components are defined as the linear combinations of the original variables that capture the most amount of variance (Jolliffe and Cadima, 2016). We look for a unit vector $\mathbf{a}_1 \in \mathbb{R}^p$ that will maximize the variance of the projection of the data onto $\mathbf{a}_1$. The projection of each observation $\mathbf{x}_j$ onto $\mathbf{a}_1$ produces the variable $y_1 = \mathbf{a}_1^{\top}\mathbf{x}_j$. The variance for this projection of the centered data is:

$$Var\left(y_1\right) = Var\left(\mathbf{a}_1^{\top}\mathbf{x}\right) = \mathbf{a}_1^{\top}\mathbf{S}\mathbf{a}_1, \tag{12}$$

because $\mathbf{S}$ is defined as the covariance matrix associated with the dataset. We want to choose $\mathbf{a}_1$ so that it maximizes $\mathbf{a}_1^{\top}\mathbf{S}\mathbf{a}_1$. This is a constrained optimization problem: maximize $\mathbf{a}^{\top}\mathbf{S}\mathbf{a}$, subject to $\mathbf{a}_1^{\top}\mathbf{a}_1 = 1$.
Using a Lagrange multiplier λ, we set $\mathcal{L}(\mathbf{a}) = \mathbf{a}^{\top}\mathbf{S}\mathbf{a} - \lambda(\mathbf{a}^{\top}\mathbf{a} - 1)$. Taking the derivative with respect to $\mathbf{a}$ and setting this to zero gives us the classical eigenvalue equation:

$$\mathbf{Sa} = \lambda \mathbf{a}. \tag{13}$$

This means that $\mathbf{a}$ must be an eigenvector of $\mathbf{S}$ while λ is its corresponding eigenvalue. The direction that maximizes variance is given by the eigenvector of the covariance matrix $\mathbf{S}$ (Jolliffe and Cadima, 2016). The solution of the first principal component $\mathbf{a}_1$ is associated with the largest eigenvalue λ_1, since $\mathbf{a}_1^\top \mathbf{S}\mathbf{a}_1 = \lambda_1$ for $\mathbf{a}_1$, which is the maximum achievable variance. Subsequent principal components are found by looking for the next orthogonal direction that captures the most remaining variance (the second largest eigenvalue). If we order the eigenvalues of $\mathbf{S}$ as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, where λ_i is the absolute value which does not change the direction, and take $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ as the corresponding eigenvectors, then $\mathbf{a}_1$ is the direction of the maximal variance and therefore the first PC (Maćkiewicz and Ratajczak, 1993). $\mathbf{a}_2$ is the direction of maximal variance orthogonal to $\mathbf{a}_1$ and PC two. This can be done up to PC p, which would explain the least amount of additional variance.

All eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$ form the orthonormal basis for the space $\mathbf{a}_i^\top \mathbf{a}_j = 0$ where $i \neq j$ and the eigenvalues $\lambda_1, \ldots, \lambda_p$ represent the variance of the data projected on the corresponding principal component directions.

PCA allows us to represent each observation in a lower-dimensional space, where the loadings of the PC's define how the original variables contribute to each axis of the new space.

5.4 Partial dependence plots

Partial dependence plots (PDP) are a visual illustration which show the marginal effects that one or two features have on the predicted outcome of a model (Friedman, 2001). With this tool, it is possible to visualize the relationship between features and the target and differentiate the impact and effect a variable has on the models prediction. Let $\mathbf{X}$ be a set of predictor variables, while $\mathbf{X}_S$ is a subset containing the variable or variables of interest. The $\mathbf{X}_C$ is the other features used in model f. The partial dependence function

is defined as:

$$PD = f_S(\mathbf{X}_S) = \mathbb{E}_{\mathbf{X}_C}[f(\mathbf{X}_S, \mathbf{X}_C)] = \int f(\mathbf{X}_S, \mathbf{X}_C) d\mathbb{P}(\mathbf{X}_C). \tag{14}$$

Partial dependence works by marginalizing the model f over the distribution of the features $\mathbf{X}_C$, resulting in a function that depends only on the features of $\mathbf{X}_S$, which show how the features in $\mathbf{X}_S$ affect the prediction. The partial functions $\hat{f}_S$ are defined as:

$$\hat{f}_S(\mathbf{X}_S) = \frac{1}{N} \sum_{i=1}^{n} f(\mathbf{X}_S, \mathbf{X}_C^{(i)}), \tag{15}$$

which gives us - for a given value of selected features S - the average marginal effect it has on the prediction. It is important to note that you assume the features in C are uncorrelated with S, which would otherwise lead to biased estimates. For categorical features, the PDP is calculated by taking an estimate for each category and replacing these for all instances of the variable in the data and averaging the predictions for each occurrence (Molnar, 2020). For the implementation, the R package from pdp Greenwell (2017) was used. Further packages that were used include packages: gridextra from Auguie (2017), pur from Wickham and Henry (2025) and tidyverse from Wickham et al. (2019). Other packages used can be found in the R repository in the electronic appendix B.

6 Results

With the data at hand, we looked at what defines a good or bad transfer (definition 1) while we want to know how to predict a good or bad transfer, based on data of the previous season before a transfer happened. For this we looked at each transfer independent from each other transfer and did not take into account data of previous seasons. First we imputed the principal components for playing styles into the data. For this each team that has played in the top five European leagues in the time frame between the 2017/18 and 2023/24 season, was analyzed based on its principal components and loaded values. Thereafter the data of all player transfers along with its associated response variable were inputed into a Random Forest model for analysis. In total eight models were fit. The models are split for each position: forward, midfield, defense and all positions; and between a binary and numeric response variable. With the fitted random forest models, the feature importance was used to determine which variables were the most influential to decide a successful transfer. Furthermore, the variable interaction in the random forest models was analyzed as well as the the features importance using partial dependence plots.

6.1 Principal Component Analysis

A feature that was non-existent in the data gathered from stathead.com was the current and previous team playing style for a player that transferred. The playing styles vary greatly for different clubs throughout the top five European leagues, which should be accounted for when discussing the predictability of a successful transfer. For this reason, playing styles where analyzed by taking into account different playing styles. For this, the team variables of 46 metrics were used to analyze different playing styles across the last six seasons of the European top 5 leagues (see Table 19, 20 and 21). This comes after a similar approach was shown by Plakias et al. (2023) to modulate playing styles across a wider range of leagues and variables. The principal component analysis here focuses on variables largely independent of of quality of play and more on the style of a particular team. Therefore variables such as goals, pass accuracy, goals conceded, amount of duels

won or similar, were excluded from this selection due to these variables being more about skill than tactics.

From the analysis there were four playing styles of teams that could be established and explain 70 % of the variance in the data. These playing styles are as follows: Possession/dominance (PC1), Gegenpressing (PC2), Defensive (PC3), Counter Attack/Wing Play (PC4). To compare particular playing styles between different teams, these four playing styles were used and compared with the principal components of the other team. These differences were calculated for each principal component and once for the total difference between all principal components combined.

In Figure 4 the different propensities of different teams in different seasons can be seen. A strong negative score indicates a propensity towards this playing style, while the variance for the former principal components is bigger and explains more variance of the data. Atletico Madrid under Diego Simeone has always played a defensive style of soccer which can be seen on this plot with them having a strong negative inclination for the defensive principal component, while also having a strong indicator for Gegenpressing. Dortmund's 2022/23 campaign bringing them into the Champions League final, has shown no stark inclination towards any particular playing style, while most teams at the top of the their league table have much of their playing style explained by the ball possession soccer (PC1) characterized by dominant teams, and also correlating with many variables which spike for good teams/teams with more possession. Eibars 2019 under José Luis Mendilibar was characterized by its strong Gegenpressing, while also being defensively weak. Sevilla on the other hand played counter attacking soccer in the 2021/22 season having less possession in their own half and crossing. Each of the PCs was introduced to the data, by the difference in playing style from the new to the old team for each transfer. In the example in 4 a player moving from Manchester City to Seville in the summer of 2022, will have a strong negative value, since it is moving from a strong possessive team to a less possessive team (the same applies the other variables, except for the total differences which is the absolute sum of all PCs)

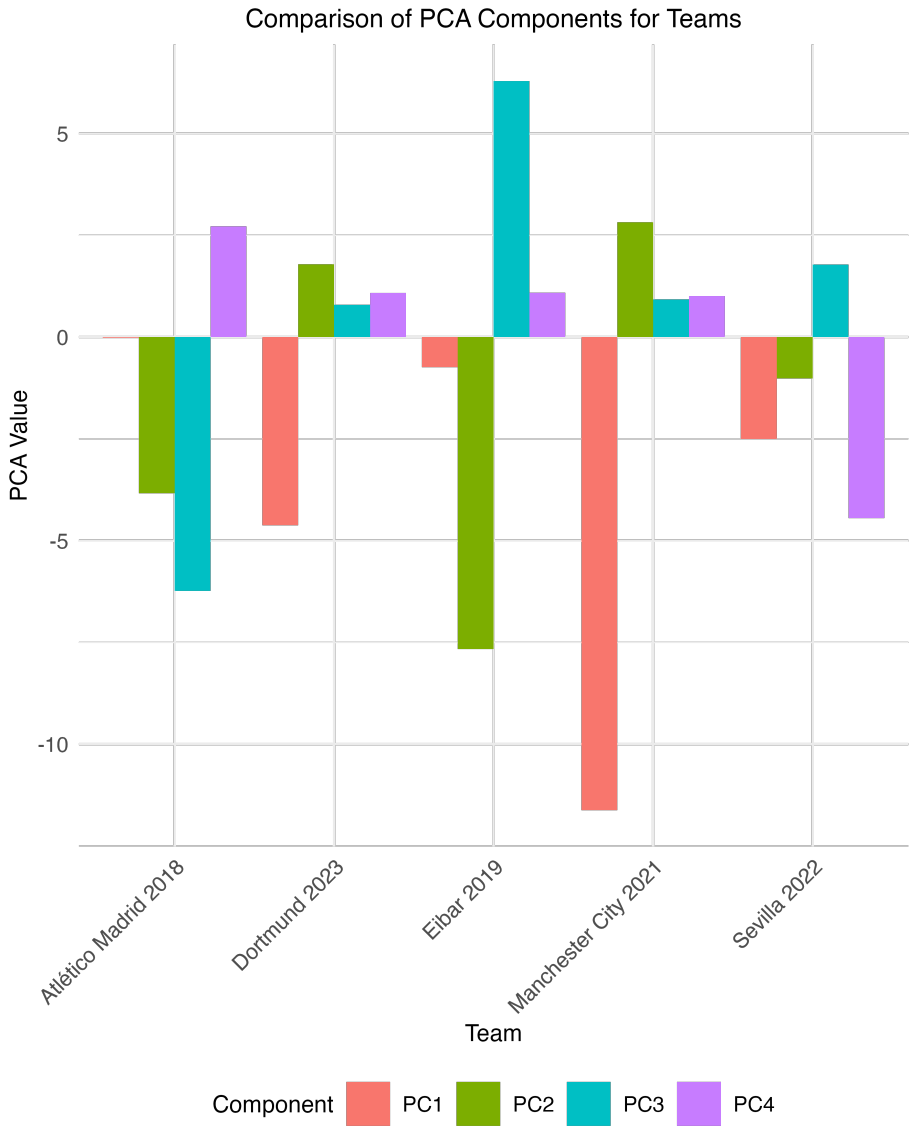

Figure 4: Differing Principal Components from Teams in the top 5 leagues

6.2 Random Forest

For each of the random forest models, hyperparameter tuning was performed. For each tree, grid search was performed for each `mtry` $\in \{50, 70, 90, 100, 110, 120, 130, 151,179\}$ and `number of trees` $\in \{50, 100, 200, 300, 400, 500, 1000\}$. For each model the hyperparameters with the best accuracy for classification tasks, and root mean square deviation error was used to determine the final model of each specification.

6.2.1 All Position Model

For the model with all positions (expect for the goalkeepers, as discussed in section 3) both models hyperparamteres were best at `mtry` = 70, while the `number of trees` for the regression model was at 500 and for the binary model at 400. The performance metrics of both models cant be compared due to the nature of both models, while the regression model hat a rmse of around 8780000, the binary model had a accuracy of about 60%. To determine what characteristics give the best indication of a successful transfer, the feature importance was used to determine the most important features. Here the variable importance of both models shared a correlation of about 75%. In Table 1 you can see the variable importance of the seven most important variables and the four least important variables based on the regression model.

Rank	**Variable**
1	Principal Component 1 Difference
2	Expected Goals Plus Per Minus Team Success
3	Principal Component 2 Difference
4	Principal Component 3 Difference
5	On-Field Minus Off-Field Team Success
6	Carries into Penalty Box
7	Plus per 90 Minutes Minus Team Success
	... (Other Variables) ...
176	Straight Corner
177	Own Goal
178	Dead Ball pass leading to goal
179	Second Yellow Card

Table 1: Top 7 and Bottom 4 Variables for Regression Model Importance Rankings

Three of the five principal component metrics are among the most important seven variables(while total difference of the PC is at rank 15 and the counter attack/wing play (PC4) parameter has rank 19). The variable importance for the binary model in Table 1 has similar characteristics, with the first principal component as the strongest variable followed by Net Expected goals for and against the Team. This is followed by the total difference in the principal components, after that coming the second and fourth principal components with Impact of player on vs Off Pitch and Expected Goals for and against

Rank	Variable
1	Principal Component 1 Difference
2	Expected Goals Plus Per Minus Team Success
3	Total Difference
4	Principal Component 2 Difference
5	Principal Component 4 Difference
6	Impact of player on vs Off Pitch
7	Expected Goals Plus per 90 Minutes Minus Team Success xG
	... (Other Variables) ...
176	Second Yellow Card
177	Straight Corner
178	Penalty Kick Attempts
179	Penalty Kicks

Table 2: Top 7 and Bottom 4 Variables for Binary Model Importance Rankings

per 90 Minutes per Team. Since both models have similar feature importance across the board (for all models), this work will focus on the binary model for the interpretation of its effects (for further information regarding the regression model please see appendix). Since random forest models do not inherently indicate if a variable has a positive or negative effect on the response variable, PDPs can be used to visualize the relationship between a variable and the model's predictions.

In Figure 5 the partial dependence plots show a selected number of variables for the binary model over all data, where the y-axis shows the incline or decrease is probability for a good transfer. Here a strong relationship between the first PC and the target variable can be seen, in this case the binary classification, if a transfer was good or not. A negative value here indicates that the team a player has moved to is less ball possessive/less dominant than the team it moved to. This can be interpreted that a transfer is more likely to be positive, if the player goes to a team that is less ball possessive/dominant than his current team. If a player therefore moves to a more ball possessive/dominant team, it is more likely that this transfer will not be good. Similar can be said about the second PC, while the third and fourth differ a lesser amount, though the interpretation of the effects stays the same for the PCs. Points per Match Team Success which is the 12th most important variable, also has strong prediction power, for a player whose team won

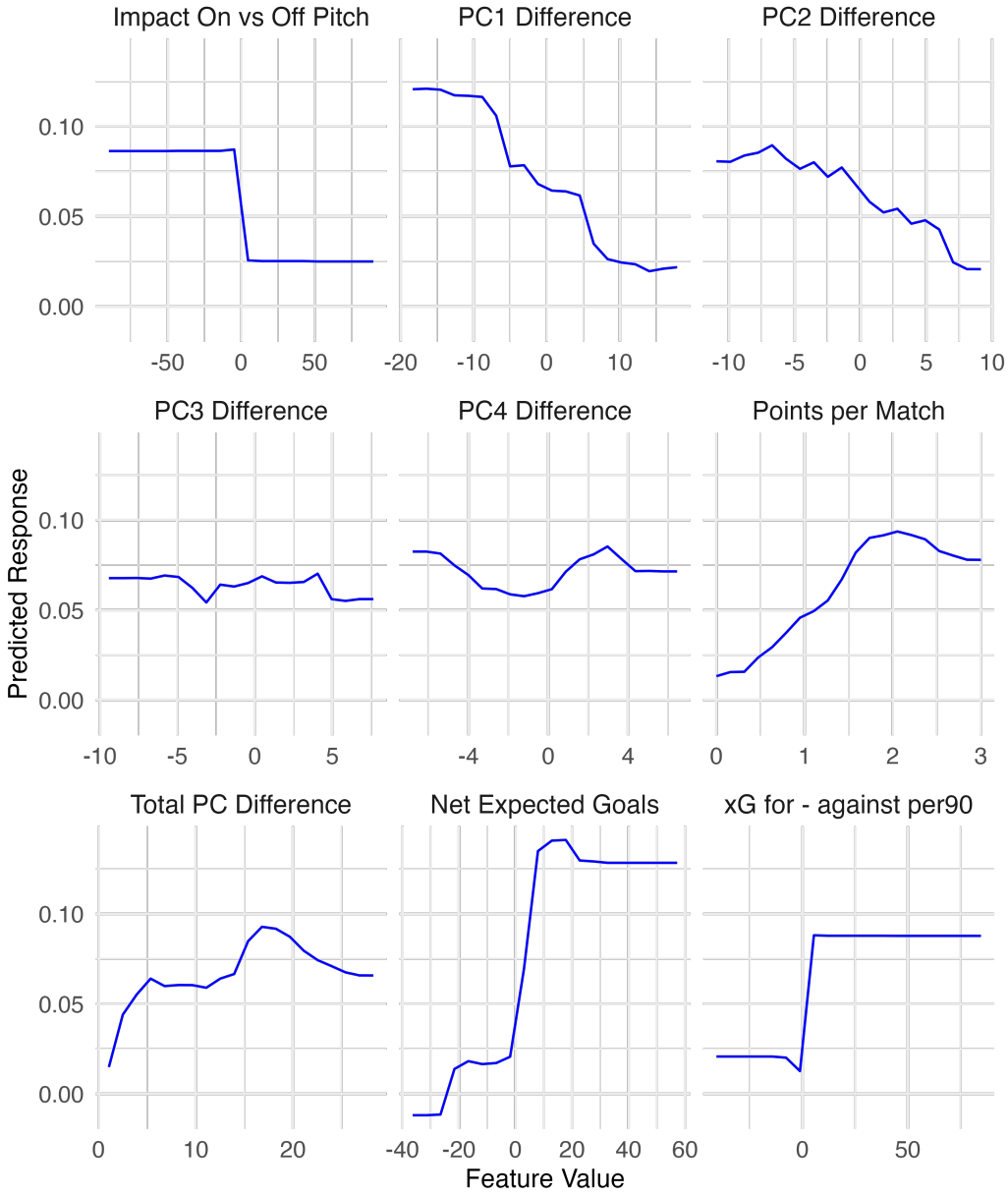

Figure 5: Partial Dependence Plot for Binary Model

on average more than 1.6 points per match the probability for a successful transfer was the highest (while he played more than 30 minutes in that match). Furthermore, the Net Expected goals for and against the Team as well as that per 90 minutes, are both among the top seven variables selected by the model and both show a stark incline in a predicted positive outcome when the expected goals for the team outweigh the expected goals from the opposing team. It is interesting to note that penalty attempts and scored

penalty's are the two least important variables, with rare events as straight corner kicks (corners that are not inswinging or outswinging) and a second yellow card also having low importance.

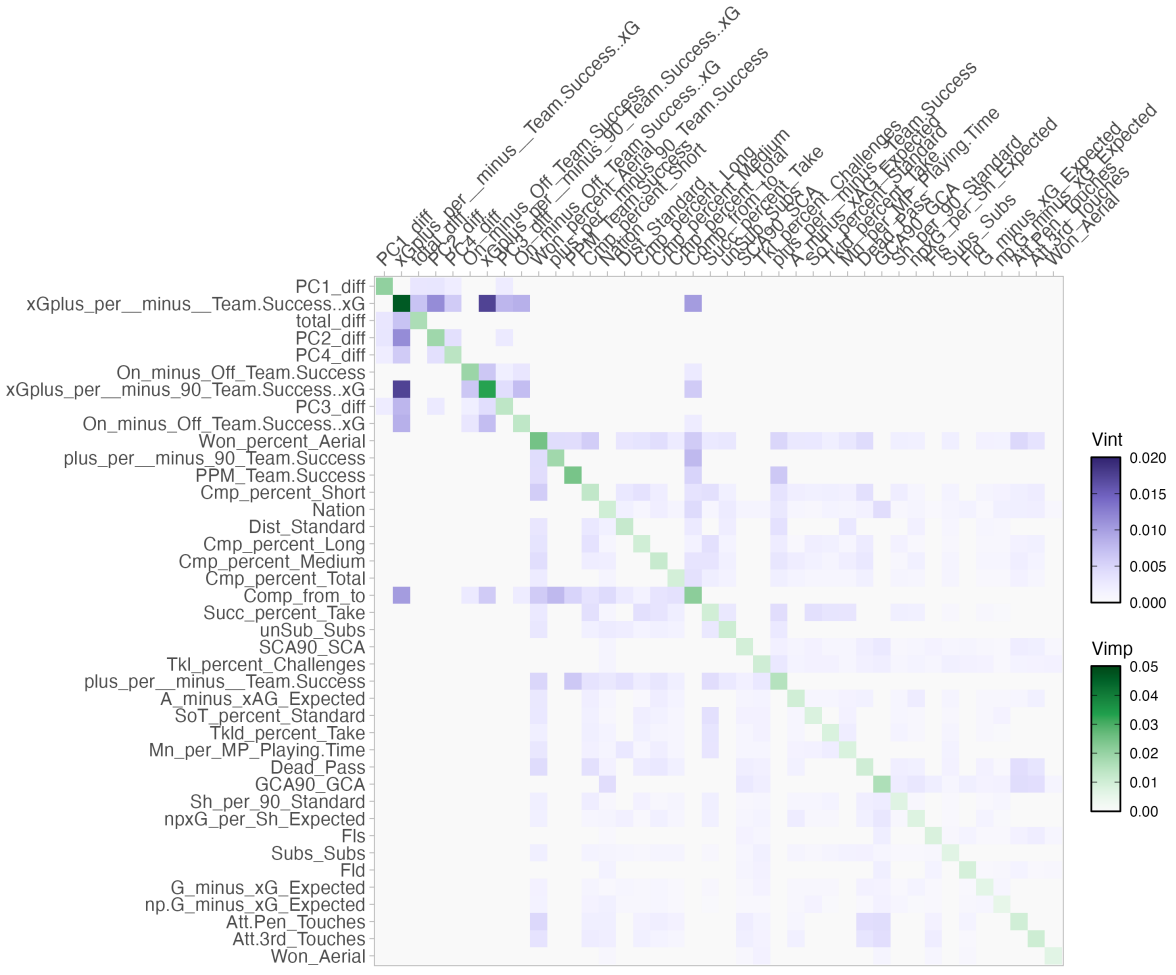

Figure 6: Interaction plot for the top 40 variables of the full binary model

Not just the features alone account for the importance of the model, but also the interactions between these features play an important role for the prediction of a transfer. Figure 6 depicts on the diagonal the ordered feature importance, whereas on the off diagonal the interaction between each variable is depicted. It is important to note that all interaction

plots show the variable importance in order of the full model, while for computational reasons the interaction plots were only fitted for the most important 40 variables, which is why the Vimp (Variable importance) is not descending in color intensity. The strongest interaction is between Net Expected goals for and against the Team and Net Expected goals for and against the Team per 90 minutes which is obvious since one variable is the other just divided by a the playing time, while both have a high variable importance. The interpretation of the effects here is not done due to the high correlation of both variables as discussed in section 5 about partial dependence plots. The second strongest interaction lies between the difference in the second principal component (Gegenpressing) and Net Expected goals for and against the Team.

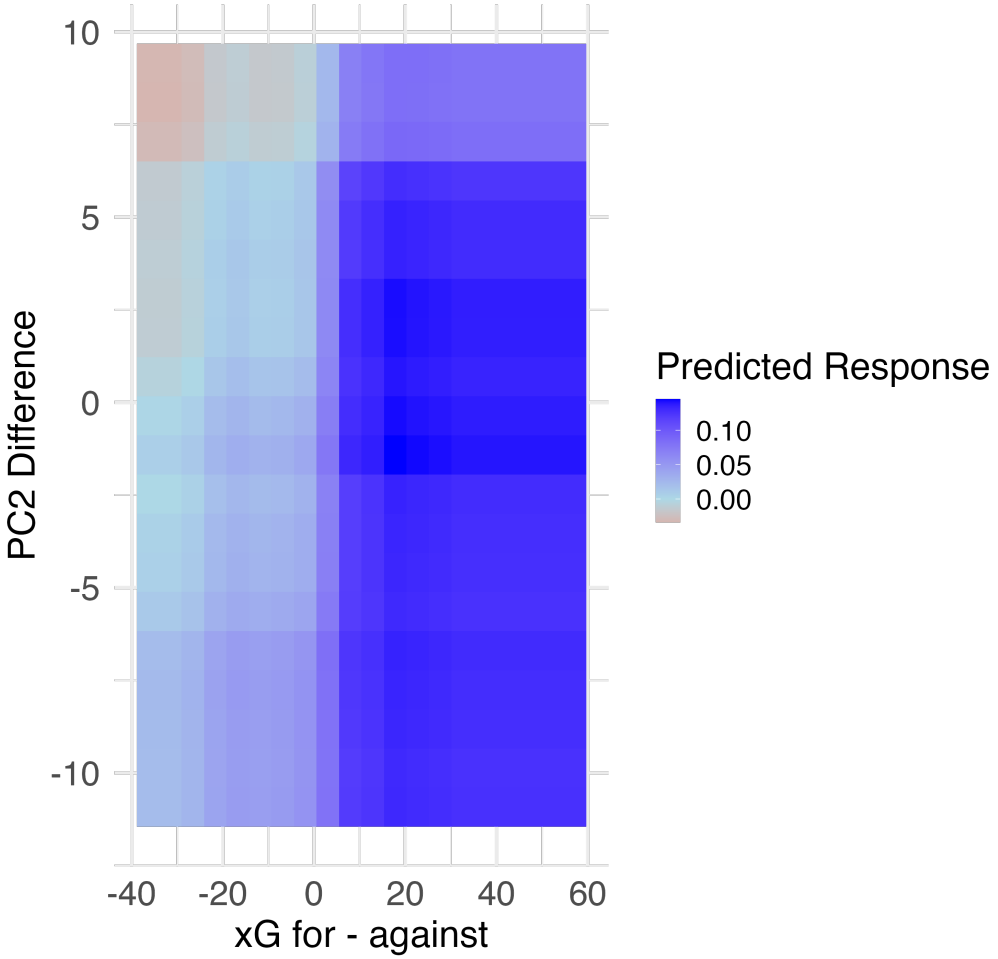

Figure 7: Interaction between Expected goals for minus against and the Gegenpressing PC for the full binary model

Here both effects combine in for a linear relationship, where high values for Net Expected goals for and against the Team and low values for Gegenpressing (PC2) result in a high prediction, whereas lower values for both variables result in low prediction values and even have a negative impact. The interaction with the third highest importance is the interaction between Net Expected goals for and against the Team and the competition a player has moves to and from. The combination of competition a player moved to and from is more important then the competition before and after alone, with both single variables not being over the top 60 variables in any other model. In Figure 8 are the

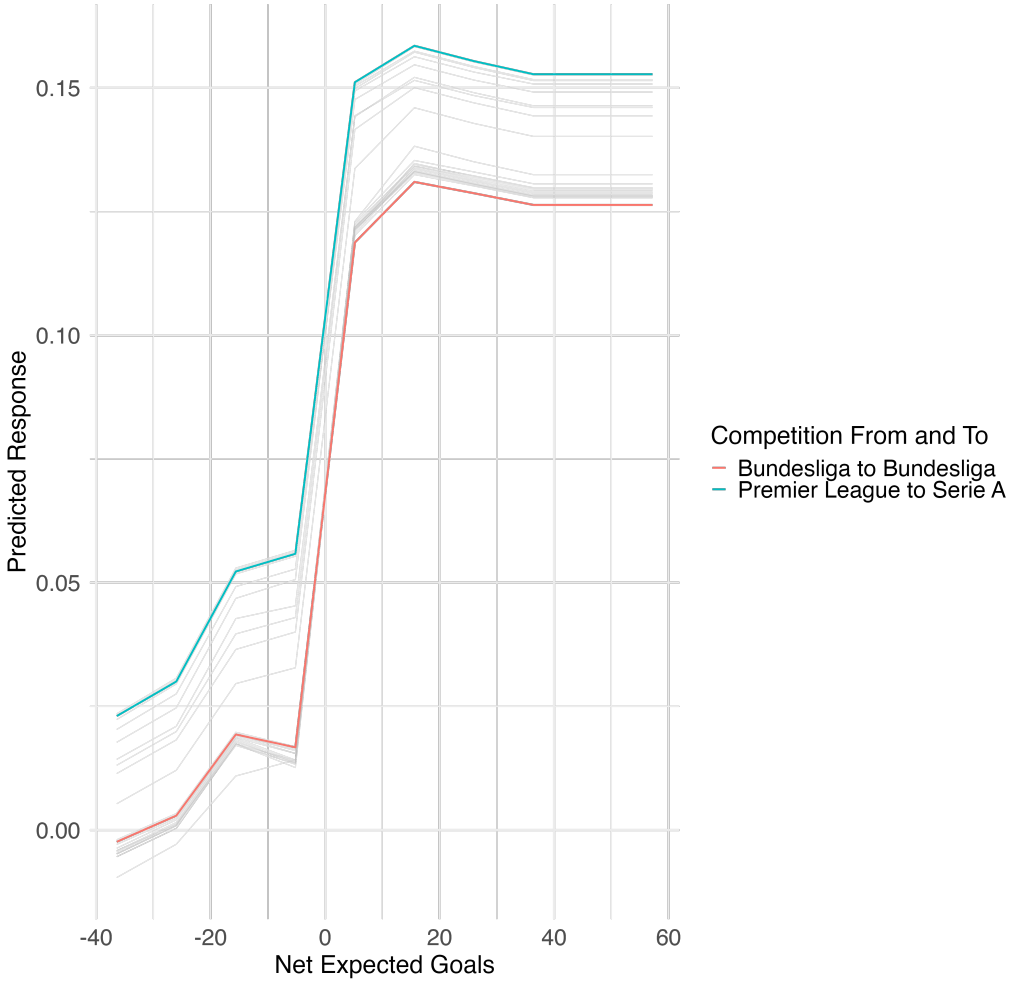

Figure 8: Interaction between Expected goals plus minus and Competition transfer for the full binary model

predictions for each combination of transfer between the top five European leagues, on the x-axis the Net Expected goals for and against the Team while on the y-axis is the

predicted probability for the partial dependence. Each combination for transfers follows a similar line dominated by the Net Expected goals for and against the Team, while each competition exhibits a random intercept, with slight changes in the curves. For most values a transfer from the Bundesliga to a Bundesliga side shows the lowest predicted response, while transfers from the EPL to Serie A show the highest of all prediction responses. The model which incorporates all outfield positions, has a variable importance of the players position on rank 83 for the regression model and on rank 63 for the binary model. Although the position of a player seems to have a relatively low effect on the prediction of the overall model, differing effects for positions can be seen for models fitted specifically for that position.

6.2.2 Forward Model

For all positions similar variables were similarly important be it for forwards, midfielders, defenders or over all positions. Still, some variables have a strongly differentiated variable importance compared to the other models. The binary forward model hyperparameters used after hyperparameter tuning was `mtry` = 110 and `number of trees` was set at 400. As visible in Table 3 the PCs also play an important role for forwards, while the successful take-on percentage is the fourth most important variable for offensive players, while this variable is the 20th most important variable for the full model, 25th for the midfielder model and 33rd for the defender model. In Figure 9 are the top six variables for the binary forward model and their partial dependencies.

The tendencies of the first three PCs are similar to the total model, where a strong negative value for the first two PCs has a positive effect on a successful transfer and positive values have has a weaker impact. The most striking difference in PCs between the model for all positions and that for forwards is the fourth PC that indicates counter attacking playing style and wing play. Whereas the values for -1 and bigger are relatively similar to the total model, values smaller than -1 have a much higher predictive probability for a successful transfer than for the model for all positions. This means that if you go from a strong wing playing or counter-attacking team, to a less counter-attacking team, it is more likely that that transfer will be successful. Furthermore, the successful take on percentage indicates,

Rank	Variable
1	Difference in Principal Component 4
2	Difference in Principal Component 1
3	Difference in Principal Component 2
4	Successful Take-On Percentage
5	Total Difference
6	Difference in Principal Component 3
7	Medium Pass Completion Percentage
8	Expected Goals Plus per Minus Team Success xG
9	Short Pass Completion Percentage
10	Points Per Match - Team Success
11	Long Pass Completion Percentage
12	Total Pass Completion Percentage
13	Nation
14	Expected Goals Plus per 90 Minutes Minus Team Success xG
15	On-Field Minus Off-Field Team Success xG
16	On-Field Minus Off-Field Team Success
17	Standard Distance
18	Plus per 90 Minutes Minus Team Success
19	Shot-Creating Actions per 90 Minutes
20	Tackles Percentage in Take-Ons

Table 3: Top 20 Variables for Forward Binary Model Importance Rankings

that a player with a high take on success rate will be more probable to be a successful transfer, while after a 40% take on success rate the effect flattens. The interaction for this model of the forwards is shown in Figure 10.

Compared to the total model in Figure 6 there are more interactions between variables across the board, while the most important interactions are between the successful take on percent and other variables. Among the most important was the interaction between the total difference in principal components and the successful take on percentage which is shown in Figure 11.

While the x-axis with the successful take-on percentage has a higher predicted value for high take-on percentages, the same cannot be said for the total difference in the total PC differences. For the total PC differences the highest prediction values occur for values between 9 and 19; towards the middle of values for the total difference, while lower values for the total difference fare worse and have a slight negative impact, compared to high

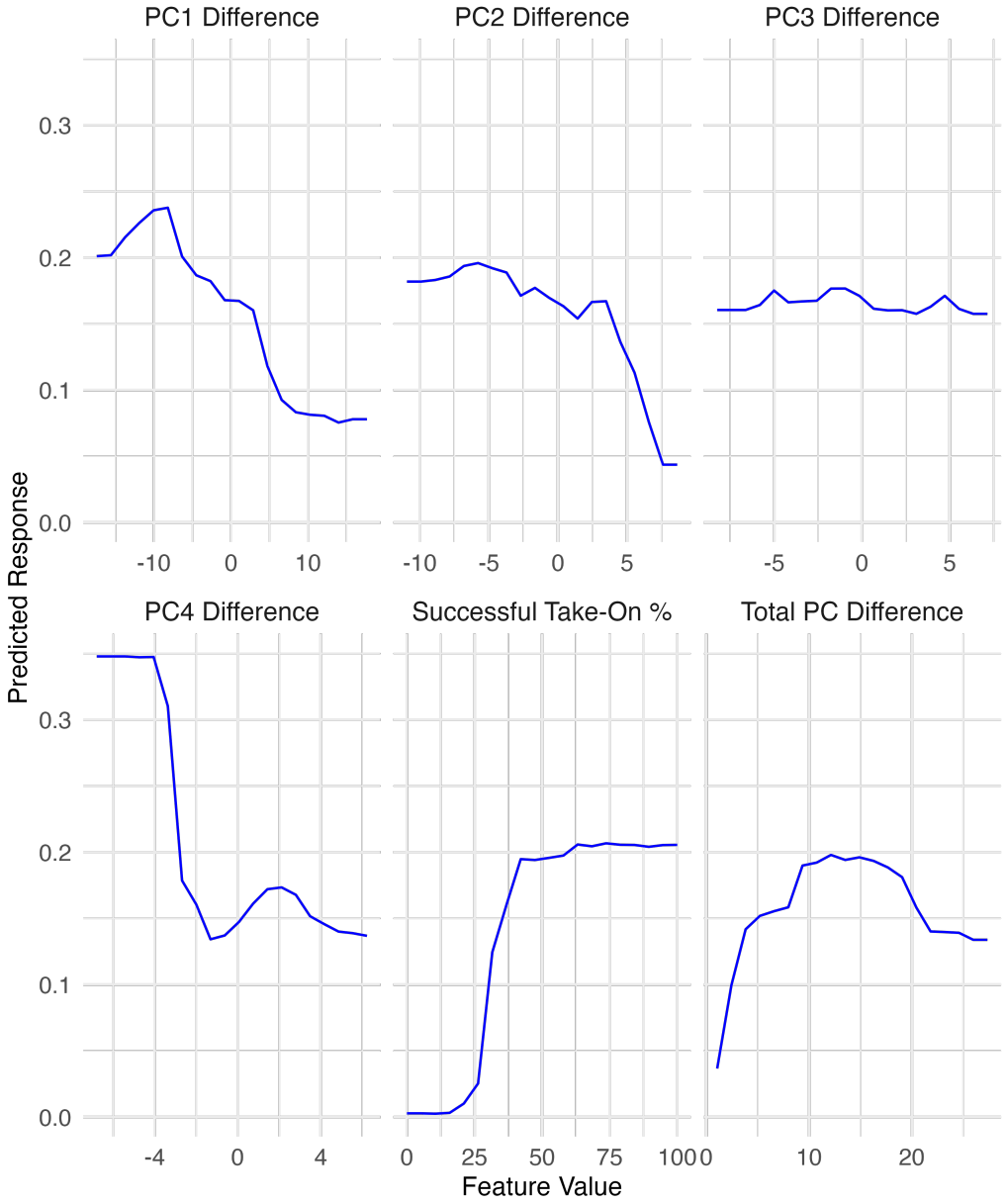

Figure 9: Partial Dependence Plot for the Binary Model for Forwards

values. The difference between the tackled during take on percentage and successful take on percentage in shown in Figure 12.

Both variables are dependent, where a high successful take on percentage correlates with a low percentage in tackled during take on, which can be seen in the plot. Even though there is a correlation between both effects, the successful take on percentage is deemed more important than the tackled during take on percentage. It is to note that due to the

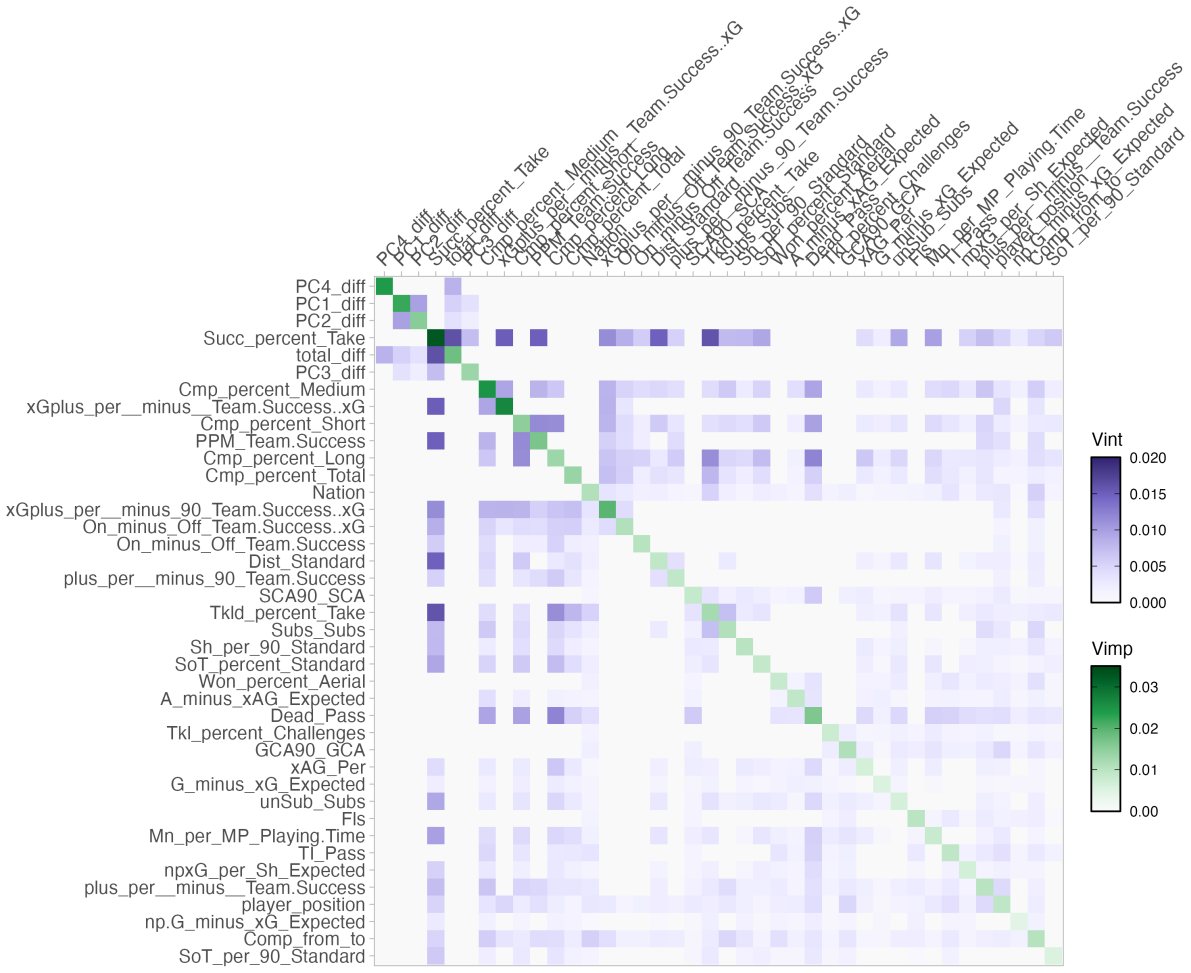

Figure 10: Interaction plot for the top 40 variables of the forward binary model

high correlation of both variables, the effect shown could be biased.

6.2.3 Midfield Model

The model for the midfielders has all PCs in its seven most important variables, while no single variable sticks out compared to the other models.

The balanced effects could be a result of midfielders taking over a variety of jobs, ranging from falling into the defensive line to pressing offensive next to the attacker.

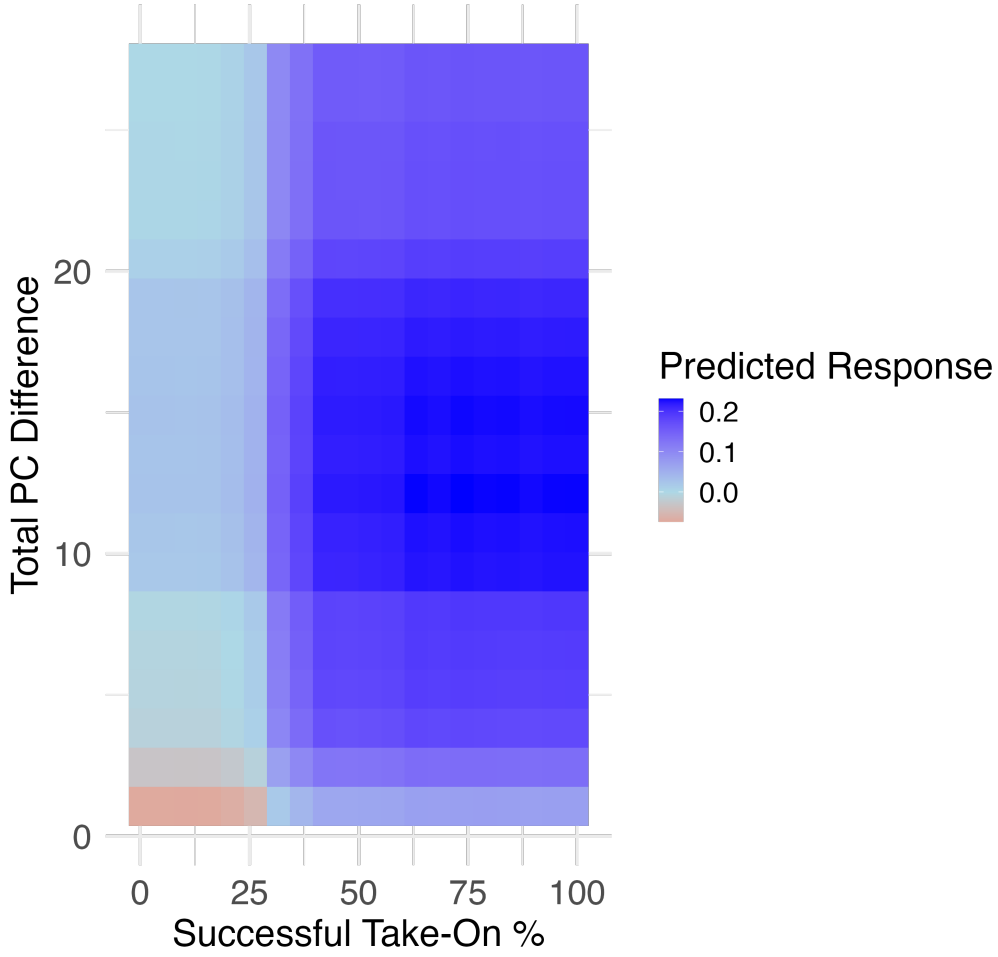

Figure 11: Interaction between the total PC differences and successful take on percentage the Binary Model for Forwards

The binary midfield model used hyperparameters with `mtry` = 100 and `number of trees` at 400, which resulted from hyperparameter tuning.

What is still obvious is the in Figure 13 depicted partial dependencies for the first and second PCs. Here both the first and second PC have similar propensities as the full and forward model with high negative values being a strong indicator for a good transfer. The third PC is contradictory to that of the defender model (Figure 17), which has a high predictive probability for high values, while that of the midfielders values strong negative values more than similar or negative values.

For the fourth PC and the total differences in PCs there are only small differences in the predicted values for both between an increase in predictive probability of 0.075 and

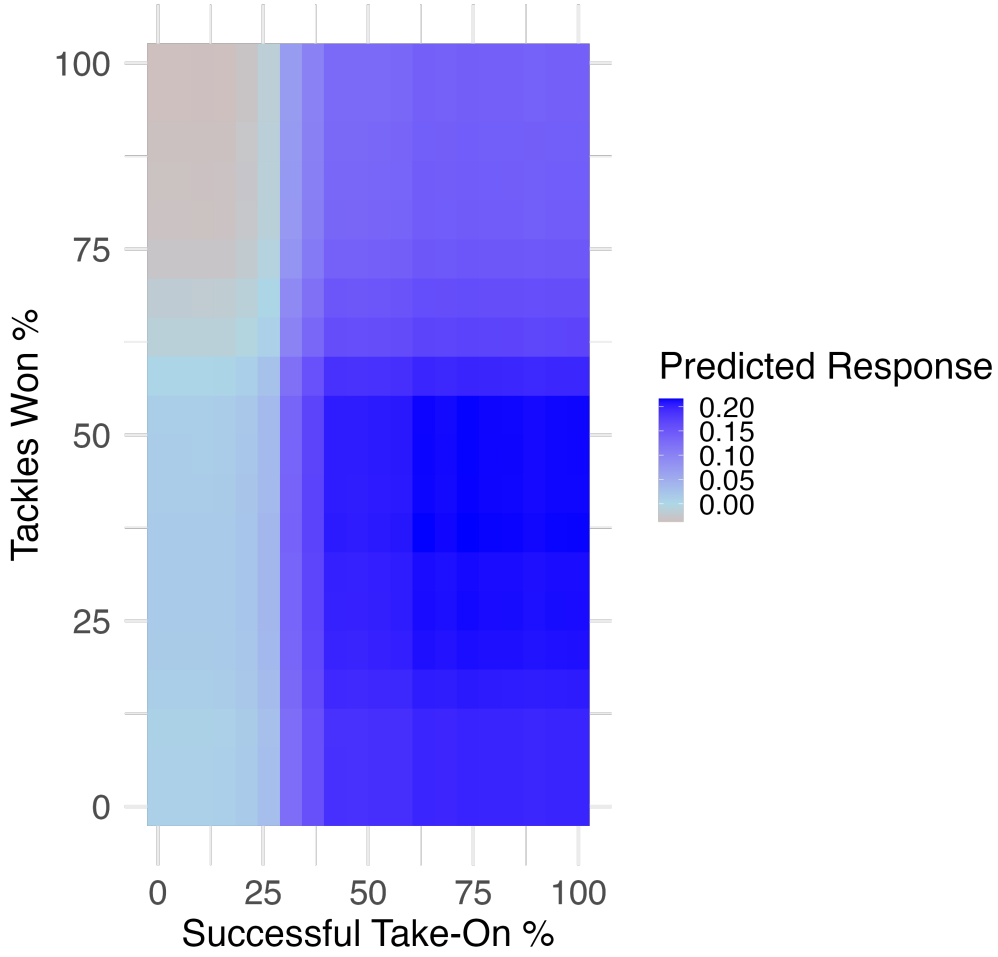

Figure 12: Interaction between the tackled during take on percentage and successful take on percentage the Binary Model for Forwards

0.12, while the fourth PC is slightly better for higher values and the total differences are best for values around 5 and 19. Expected team successes for and against per 90 has the same effect as for midfielders and the model with all positions. The interactions for the midfielder model in Figure 14 show that interactions among the most important variable are not as important, as the interactions among variables that are not as important to the model alone.

The most important interaction between variables is that between dead passes which are passes from a standard position such as a free kick, goal kick, corner or in special cases a penalty and the assists per 90 minutes. Even though both variables have a single variable importance of 24 and 34 respectively. For the partial dependence plot between dead

Rank	Variable
1	Difference in Principal Component 1
2	Difference in Principal Component 2
3	Total Difference
4	Difference in Principal Component 3
5	On-Field Minus Off-Field Team Success
6	Expected Goals Plus per 90 Minutes Minus Team Success xG
7	Difference in Principal Component 4
8	Won Percentage in Aerial Duels
9	Nation
10	Points Per Match - Team Success
11	Short Pass Completion Percentage
12	Expected Goals Plus per Minus Team Success xG
13	Plus per 90 Minutes Minus Team Success
14	Long Pass Completion Percentage
15	Medium Pass Completion Percentage
16	Assisted Minus Expected Assisted Goals
17	Completed Passes from to
18	On-Field Minus Off-Field Team Success xG
19	Standard Distance
20	Shot-Creating Actions per 90 Minutes

Table 4: Top 20 Variables for midfielder Binary Model Importance Rankings

passes and assists per 90 minutes in Figure 15 there is a strong distinction in between dead passes over and under about 20.

This distinction intuitively makes sense, since players who play these types of passes are often the same, which means that players taking free kicks or corners are the same players; you either have a lot of dead passes or none. For dead passes greater than this threshold, there is a higher predicted response than values below. There is also a distinction between more and less than 0.7 assists per 90 minutes, while less assists are connected to a higher predicted response. The sweet spot for for assists lies between 0.1 and 0.7 per 90 minutes for midfielders and about 50 to 250 dead passes. The second most important interaction is depicted in Figure 16 between dead passes and won aerial duals, which is similar in nature to Figure 15, with dead passes exhibiting a threshold for lower values. Won ariel duals as its highest predictive response at around 25 percent, with effects staying relatively constant throughout dead passes over 20.

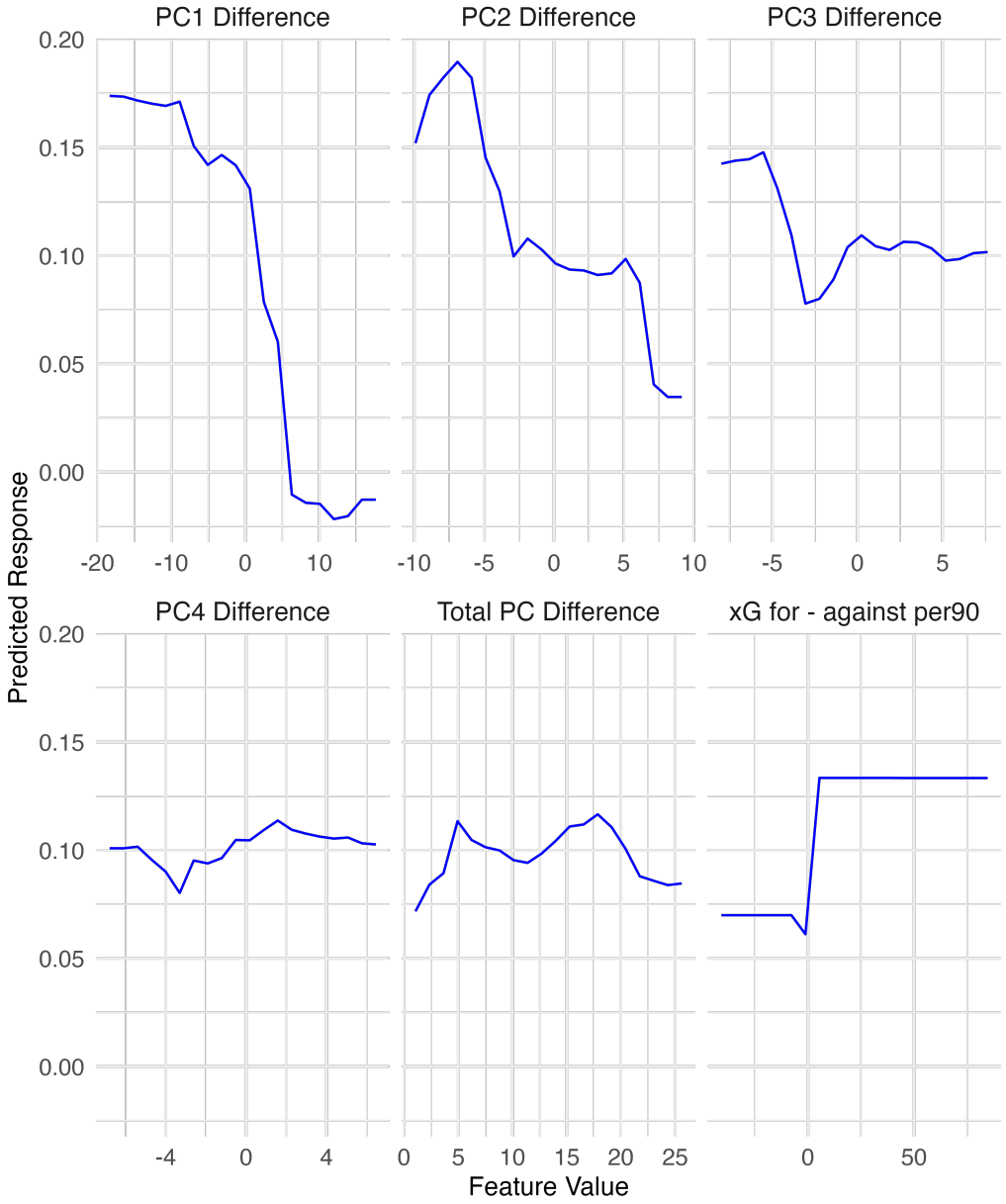

Figure 13: Partial Dependence Plot for the Binary Model for Midfielders

6.2.4 Defender Model

The defender model can be seen as a model of multiple positions too. The binary defender models hyperparameters used in the final model were `mtry` = 120 and `number of trees` = 300. Since central defenders and fullbacks often have fundamentally different functions for ones team. Although there is a discrepancy between the positions, in the model, the players position only had a variable importance of 130. This could be explained by the

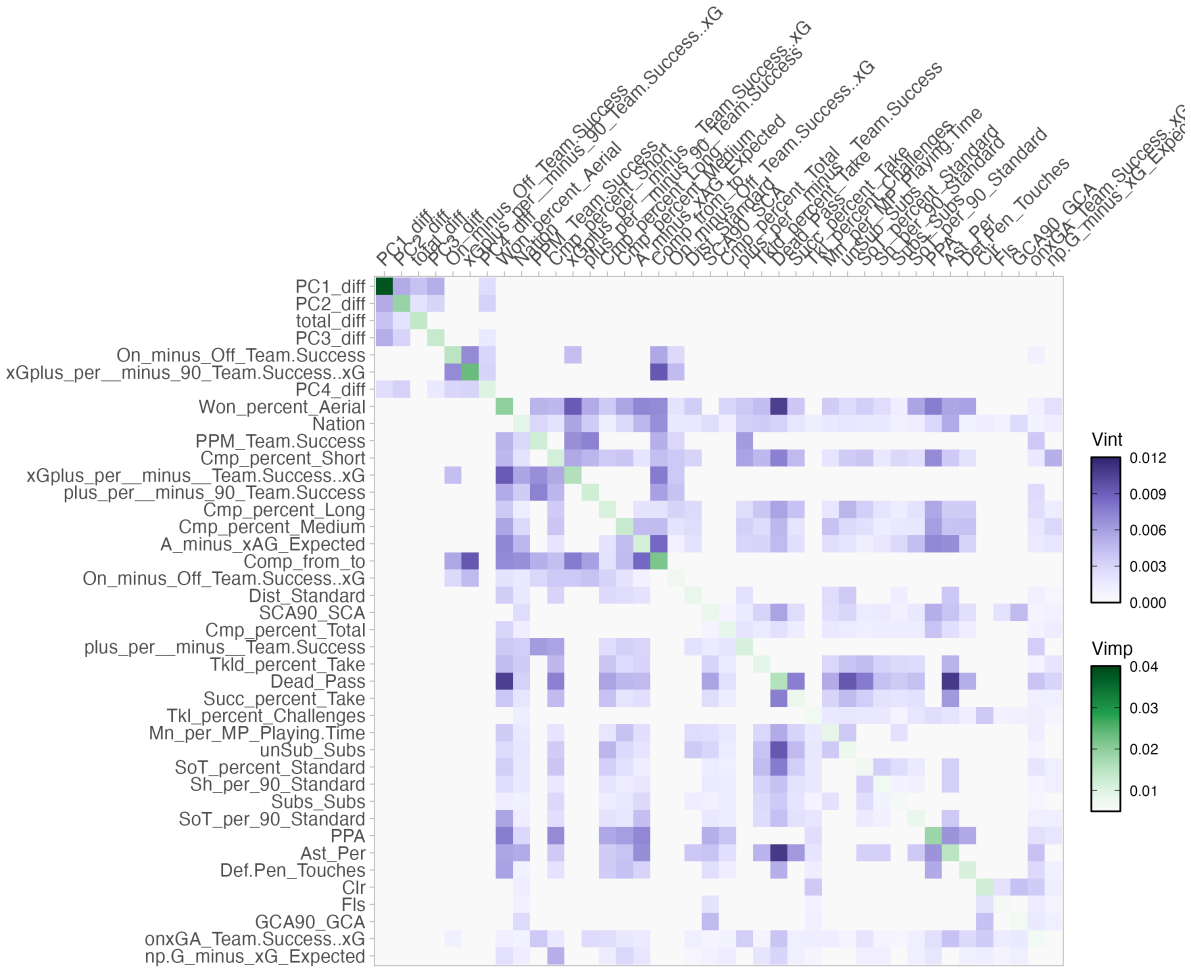

Figure 14: Interaction plot for the top 40 variables of the midfielder binary model

tactical objective of defenders as a whole, having to defend more often than any other positions. The variable importance for defenders shown in Table 5, is congruent with that of the other models, with all five PC variables being among the most important seven. While Expected Goals Plus Per Minus Team Success is also among the most important variables, the percentage of won aerial duels is the second most important variable for defenders. This comes as no surprise, since defenders often come forward for corners or standard situations and also have to defend against long balls throughout the game.

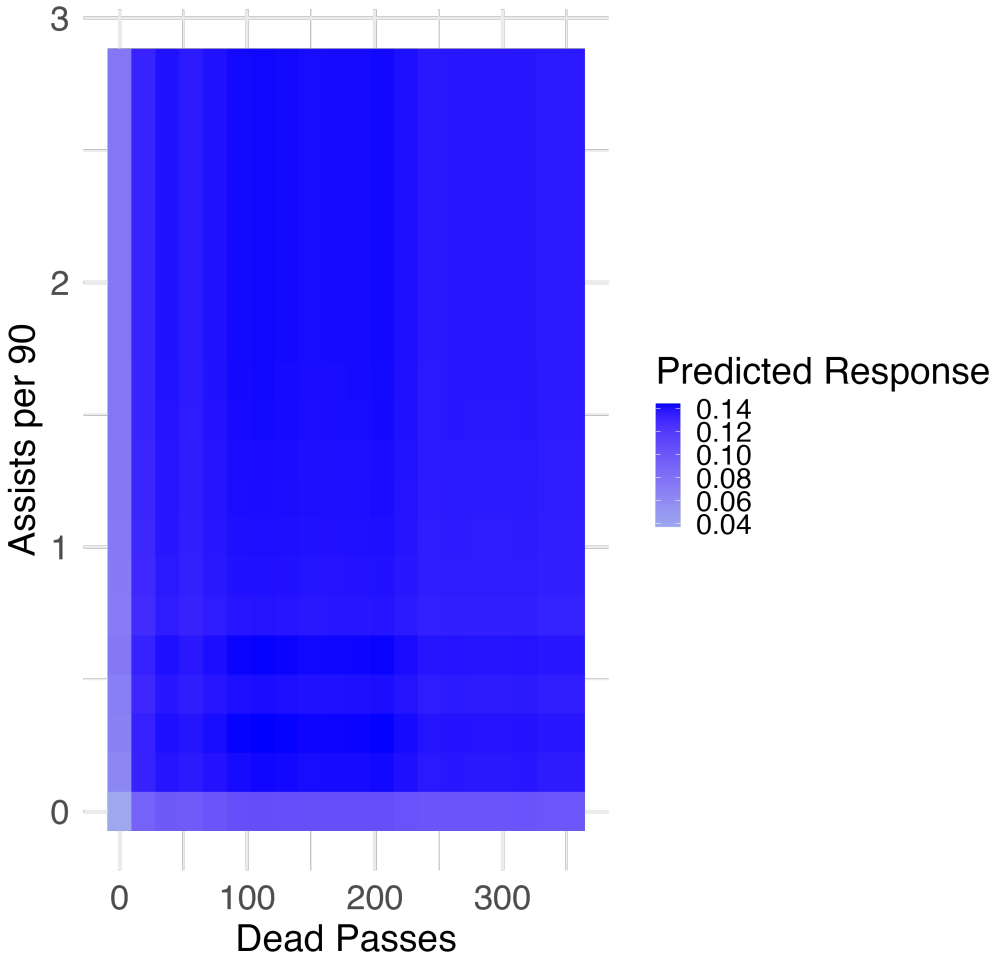

Figure 15: Interaction between Dead passes and Assists per game for the Binary Model for Midfielders

As for the PCs there is a difference in their propensities compared to other models. In Figure 17 the first and second PCs have stark differing proclivities. Here, the first PC sees the same behavior as the other models for values smaller than one, but goes back up until about 5, which after slowly sinks again and has a negative predictability. Also for the second PC you can see a totally different behavior, with values close to similar playing style for Gegenpressing are the best predictor for a successful transfer, while positive values are worse than negative values. The third PC also differs from the other models, in that higher values indicate a better transfer than lower values, where the other models had a relatively constant effect, with the effects being almost only negative. For defenders, variables and PDPs often indicate negative values, which could be a result of defenders

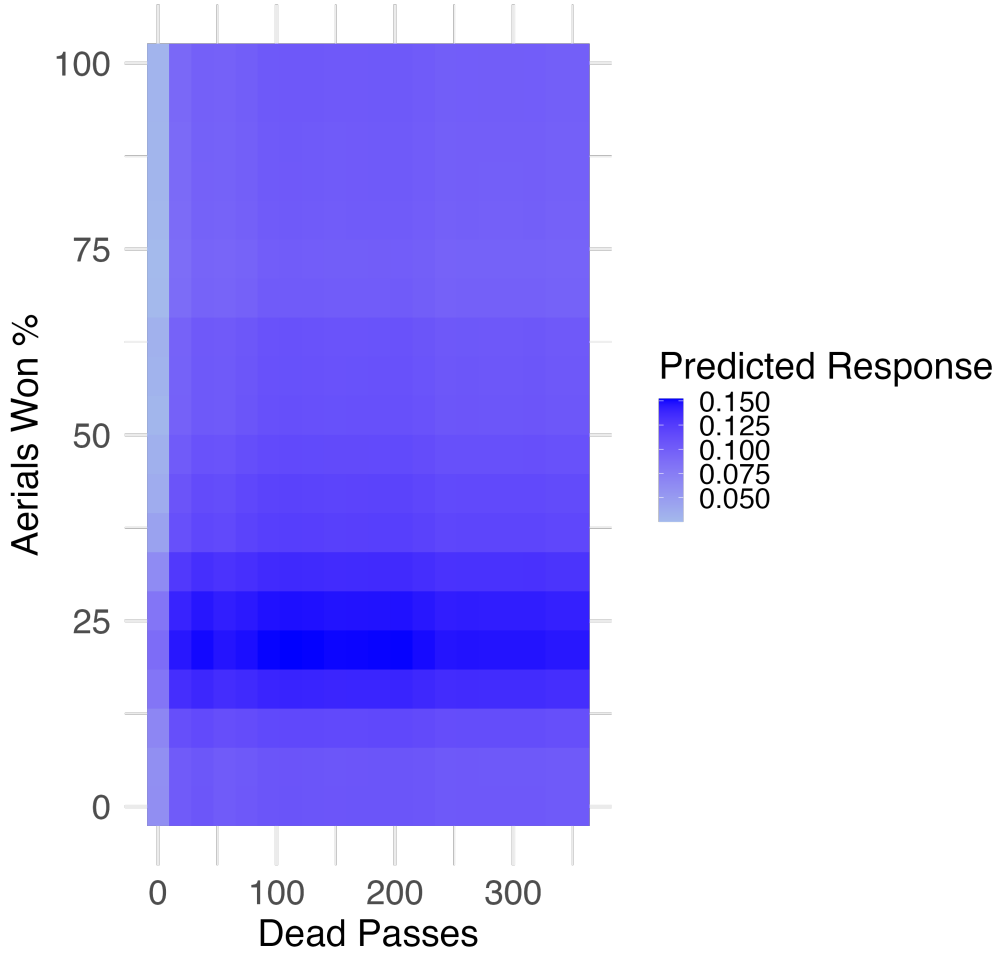

Figure 16: Interaction between Dead passes and Won Aerial Duals for the Binary Model for Midfielders

on average having a different set of characteristic than midfielders and forwards, also seen in the differing market values with defenders having on average a lower market value than their attacking counterparts. The most interesting PDP is the percentage of won aerial battles, which has high predictive probability for low percentages, while negative predictive probability for high percentages. All this information also translates into the variable interactions of the model in Figure 18. In the top 40 variables for the binary model for the defenders, there are many interactions between the variables, while the most important interactions are almost exclusively bound to the percentage of won aerial duels. The most important of which were the with Net Expected goals for and against the Team per 90 followed by Net Expected goals for and against the Team which is the

Rank	Variable
1	Difference in Principal Component 2
2	Won Percentage in Aerial Duels
3	Difference in Principal Component 4
4	Total Difference
5	Expected Goals Plus per Minus Team Success xG
6	Difference in Principal Component 1
7	Difference in Principal Component 3
8	On-Field Minus Off-Field Team Success xG
9	On-Field Minus Off-Field Team Success
10	Expected Goals Plus per 90 Minutes Minus Team Success xG
11	Unsuccessful Substitutions
12	Completed Passes from to
13	Short Pass Completion Percentage
14	Medium Pass Completion Percentage
15	Points Per Match - Team Success
16	Long Pass Completion Percentage
17	Nation
18	Total Pass Completion Percentage
19	Tackle Percentage in Challenges
20	Fouls Committed

Table 5: Top 20 Variables for Defenders Binary Model Importance Rankings

un-normed variable, not based on playing time. For Net Expected goals for and against the Team the results are depicted in Figure 19.

A low percentage in aerial battles won and positive expected goals for the team have the best predicted response while the opposite in both values result in a negative prediction for the response. The third most important interaction in this model is the Impact on vs off Pitch for a team which is similar to the Net Expected goals for and against the Team, though how good a team performed with or without him on, with the addition that penalties were not counted in the former.

6.2.5 Similarities between all Models

Each model has different propensities towards different characteristics. Nevertheless, there are similarities across all models irrespective of position or other factors. For one the completion percentage of passes for short, medium and long, as well as total passes were all

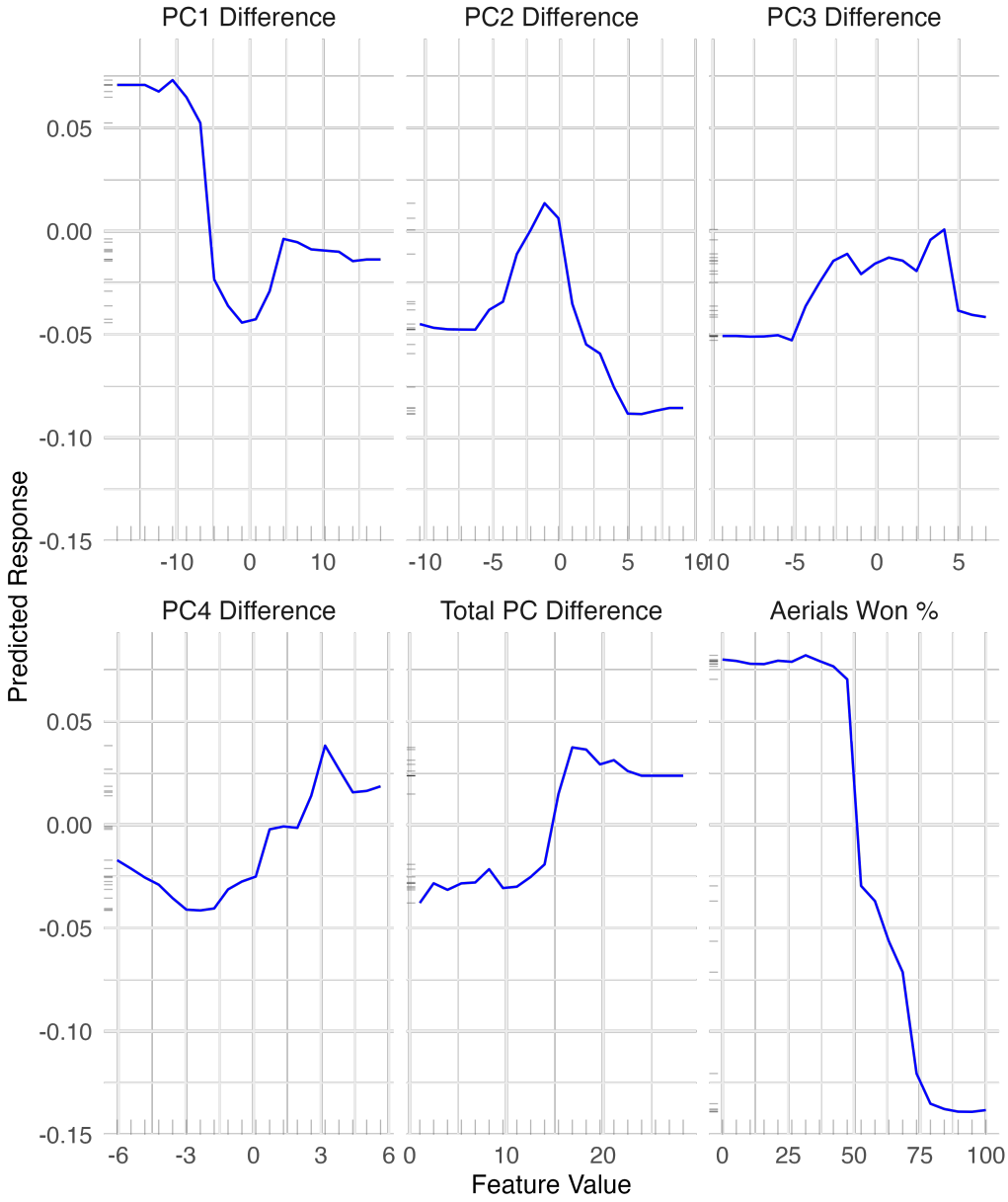

Figure 17: Partial Dependence Plot for the Binary Model for Defenders

in the top 20 most important variables irrespective of the model. The variable importance for passes attempted and completed for these categories were all above 60, indicating that while passes are important for a successful transfer, volume alone is not important for the prediction - but quality is. An additional variable that is rated highly in every model is nation. The nation seems to have an important impact on the models at hand with Nation also being among the top 20 most important variables in every model. Even

Figure 18: Interaction plot for the top 40 variables of the defender binary model

though the objective of a soccer match is to score as many goals as possible, goals are among the least important variables for all models with it not ranking above 171 of 179 among all model's variable importance. Although the goals variable is not important for any model, goals per 90 minutes is relatively important for attackers, being on rank 45 (not ranking any higher than 112 for any other model). Another characteristic for all models, are that the per 90 minutes statistics, are more important to the model than the absolute values of those statistics. With over 10 values in the dataset being absolute as

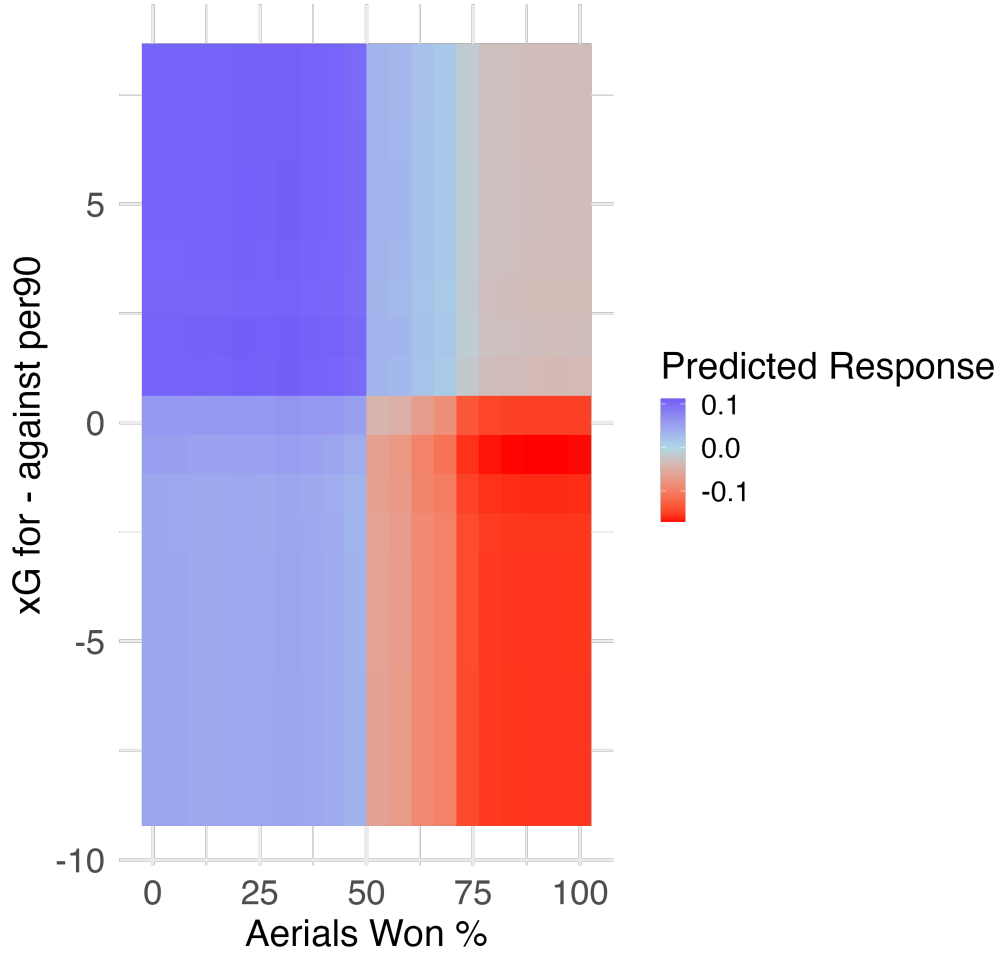

Figure 19: Interaction between the Net Expected goals for and against the Team and Won Aerial Duals for the Binary Model for Defenders

well as per 90 statistics, all per 90 variables have shown a higher variable importance, with the exception being Net Expected goals for and against the Team. In addition, expected values are valued having a higher feature importance than the absolute value of a statistic such as expected goals and goals. This can be expected since expected statistics reduce the effect of randomness and luck, while accounting for the quality rather than just results (Mead et al., 2023).

7 Conclusion

This work presented a new approach to predict successful transfers in the top five European soccer leagues. The models showed varying success in the prediction of successful transfers, yet provided valuable insights into the contributing factors for a successful transfer. Team playing styles aggregated from descriptive performance metrics analyzed by principal component analysis provided valuable insight into a transfer prediction. Each position-based model and the complete model vary slightly in its categorization of important variables. All models selected the imported principal components among the most important variables, while also selecting team based performance metrics higher on average than individual performance metrics. This may indicate that a teams performance when a player is on the field has an impact that cannot be measured by standard data collection, but may be seen in Player Efficiency Ratings such as proposed by Deshpande and Jensen (2016) for the NBA. Furthermore, relatively rare events such as own goals, won penalty kicks, second yellow cards, standard goals, etc. have little to no significance in determining a successful transfer among all models. While this work has focused on the binary model primarily, regression models provided only limited insights into incremental changes in value, struggling to adapt to outliers. For further information see the appendix for partial dependence plots of the regression model. The results for the regression-based model indicated similar variable importance patterns, with a stark emphasis on team level performance metrics, despite being outnumbered by player-based performance metrics.
This work can be used to improve the understanding and decision making for teams, by not only focusing on player metrics do decern a good or promising player, but also by analyzing the team and system surrounding the player. The result can lead to a more team building approach, focusing less on the individual player, but rather on his capabilities to function in a similar environment. This concept may be especially resourceful for teams with a lower budget for smaller expenditures.
There are some limitations to this method though. Although the models insights can be used to analyze past and future transfers, the prediction strength of the overall models are weak. Another problematic aspect is the handling of variables such as the age and market value of a player. For all models the variables for age and market value where

deliberately left out. This results in the model not having all data at hand to distinguish a successful transfer for the definition provided. Since the definition of a successful transfer in this work is heavily dependent on both variables, these were left out. Both variables - if in the models - were highly important and showed that young age and high market value have a positive impact on a successful transfer by definition of this work. Lastly, the definition of a successful transfer is based solely on the market value of a player and does not incorporate other criteria teams might have for a successful transfer. It also has to be stressed that this work does not take into account causal relationships. The characteristics provided for a successful transfer only indicate a associative relationship between both the action and a successful transfer.

For future research it would be interesting to look at different definitions of a successful transfer and if the same team-based metrics and principal components are as important for that definition of success. Another application could be made for women's soccer leagues, while data could also be expanded to more leagues than just the top five European leagues for men. Further research could be done using data not just from the preceding season, but also seasons predating that.

List of Figures

1 Median value per age, cut off at values -10 Million and 10 Million, for better visualization, representing about 90% of of transfers 6
2 A simple Classification and Regression Tree (CART) for regression. 9
3 A simple Classification and Regression Tree (CART) for classification. . . . 10
4 Differing Principal Components from Teams in the top 5 leagues 18
5 Partial Dependence Plot for Binary Model 21
6 Interaction plot for the top 40 variables of the full binary model 22
7 Interaction between Expected goals for minus against and the Gegenpressing PC for the full binary model 23
8 Interaction between Expected goals plus minus and Competition transfer for the full binary model 24
9 Partial Dependence Plot for the Binary Model for Forwards 27
10 Interaction plot for the top 40 variables of the forward binary model 28
11 Interaction between the total PC differences and successful take on percentage the Binary Model for Forwards 29
12 Interaction between the tackled during take on percentage and successful take on percentage the Binary Model for Forwards 30
13 Partial Dependence Plot for the Binary Model for Midfielders 32
14 Interaction plot for the top 40 variables of the midfielder binary model . . 33
15 Interaction between Dead passes and Assists per game for the Binary Model for Midfielders 34
16 Interaction between Dead passes and Won Aerial Duals for the Binary Model for Midfielders 35
17 Partial Dependence Plot for the Binary Model for Defenders 37
18 Interaction plot for the top 40 variables of the defender binary model . . . 38
19 Interaction between the Net Expected goals for and against the Team and Won Aerial Duals for the Binary Model for Defenders 39

20 PDP for the top 20 variables for the the binary model for all positions (excluding categorical variables such as nation and the competition transferred to and from) XXII
21 PDP for the top 20 variables for the the binary model for forwards (excluding the categorical variable: nation) XXIII
22 PDP for the top 20 variables for the the binary model for midfielders (excluding categorical variables such as nation and the competition transferred to and from) XXIV
23 PDP for the top 20 variables for the the binary model for defenders (excluding categorical variables such as nation and the competition transferred to and from) XXV
24 PDP for the top 20 variables for the the regression model for all positions (excluding the categorical variable: nation) XXVI
25 PDP for the top 20 variables for the the regression model for forwards . . . XXVII
26 PDP for the top 20 variables for the the regression model for midfielders (excluding the categorical variable: nation) XXVIII
27 PDP for the top 20 variables for the the regression model for defenders . . XXIX

List of Tables

1 Top 7 and Bottom 4 Variables for Regression Model Importance Rankings 19
2 Top 7 and Bottom 4 Variables for Binary Model Importance Rankings . . . 20
3 Top 20 Variables for Forward Binary Model Importance Rankings 26
4 Top 20 Variables for midfielder Binary Model Importance Rankings 31
5 Top 20 Variables for Defenders Binary Model Importance Rankings 36
6 Descriptions of FBref the Standard Variables 1 IX
7 Descriptions of FBref the Standard Variables 2 X
8 Descriptions of Selected FBref Shooting Variables XI
9 Descriptions of Selected FBref Passing Variables XII
10 Descriptions of Selected FBref advanced Passing Variables XIII
11 Descriptions of Selected FBref Shot and Goal Creation Variables 1 XIV
12 Descriptions of Selected FBref Shot and Goal Creation Variables 2 XV
13 Descriptions of Selected FBref Defensive Variables XVI
14 Descriptions of Selected FBref Possession Variables 1 XVII
15 Descriptions of Selected FBref Possession Variables 2 XVIII
16 Descriptions of Selected FBref Playing Time and Team Success Variables . XIX
17 Descriptions of Selected FBref Miscellaneous Variables XX
18 Description of Variables in the Transfermarkt Dataset XXI
19 Passing and Progression Statistics. XXX
20 Shooting and Defensive Contributions. XXXI
21 Touches, Aerial Duels, and Fouls. XXXI

A Appendix

A.1 Variables imported from FBref and Transfermarkt

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
MP_Playing	Matches Played; the total number of matches in which the player appeared.
Starts_Playing	The number of matches the player started.
Min_Playing	Total minutes played by the player throughout the season.
Mins_Per_90_Playing	Minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
Gls	Goals scored by the player.
Ast	Assists made by the player; passes leading directly to a goal.
G+A	Combined total of goals and assists by the player.
G_minus_PK	Goals scored excluding penalty kicks.
PK	Penalty kicks successfully converted by the player.
PKatt	Penalty kicks attempted by the player.
CrdY	Yellow cards received by the player.
CrdR	Red cards received by the player.
xG_Expected	Expected Goals; a metric that estimates the likelihood of a goal being scored from a particular shot based on various factors.
npxG_Expected	Non-Penalty Expected Goals; expected goals excluding those from penalty shots.
xAG_Expected	Expected Assists; the expected number of assists based on the quality of passes leading to shots.
npxG+xAG_Expected	Sum of non-penalty expected goals and expected assists.
PrgC_Progression	Progressive Carries; number of times the player carried the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any carry into the penalty area.
PrgP_Progression	Progressive Passes; number of completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area.
PrgR_Progression	Progressive Passes Received; number of times the player received a pass that moves the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any pass received in the penalty area.

Table 6: Descriptions of FBref the Standard Variables 1

Gls_Per	Goals per 90 minutes; average number of goals scored by the player per 90 minutes played.
Ast_Per	Assists per 90 minutes; average number of assists made by the player per 90 minutes played.
G+A_Per	Combined goals and assists per 90 minutes.
G_minus_PK_Per	Goals excluding penalties per 90 minutes.
G+A_minus_PK_Per	Combined goals and assists, excluding penalty goals, per 90 minutes.
Variable	**Description**
xG_Per	Expected goals per 90 minutes.
xAG_Per	Expected assists per 90 minutes.
xG+xAG_Per	Combined expected goals and expected assists per 90 minutes.
npxG_Per	Non-penalty expected goals per 90 minutes.
npxG+xAG_Per	Combined non-penalty expected goals and expected assists per 90 minutes.
Url	The specific URL link to the player's detailed statistics page on FBref.

Table 7: Descriptions of FBref the Standard Variables 2

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
Mins_Per_90	Minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
Gls_Standard	Goals scored by the player in standard play (excluding penalty kicks).
Sh_Standard	Total number of shots taken by the player.
SoT_Standard	Shots on target; number of shots that were on target.
SoT_percent_Standard	Percentage of shots that were on target.
Sh_per_90_Standard	Average number of shots taken by the player per 90 minutes played.
SoT_per_90_Standard	Average number of shots on target by the player per 90 minutes played.
G_per_Sh_Standard	Goals per shot; the ratio of goals scored to total shots taken.
G_per_SoT_Standard	Goals per shot on target; the ratio of goals scored to shots on target.
Dist_Standard	Average distance (in yards) from which the player takes shots.
FK_Standard	Number of goals scored from direct free-kicks.
PK_Standard	Number of penalty kicks successfully converted by the player.
PKatt_Standard	Number of penalty kicks attempted by the player.
xG_Expected	Expected Goals; a metric that estimates the likelihood of a goal being scored from a particular shot based on various factors.
npxG_Expected	Non-Penalty Expected Goals; expected goals excluding those from penalty shots.
npxG_per_Sh_Expected	Non-Penalty Expected Goals per shot; average xG value per non-penalty shot.
G_minus_xG_Expected	Difference between actual goals scored and expected goals.
np:G_minus_xG_Expected	Difference between actual non-penalty goals scored and non-penalty expected goals.
Url	The specific URL link to the player's detailed statistics page on FBref.

Table 8: Descriptions of Selected FBref Shooting Variables

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
Mins_Per_90	Average minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
Cmp_Total	Total number of completed passes by the player.
Att_Total	Total number of passes attempted by the player.
Cmp_percent_Total	Percentage of total passes completed by the player.
TotDist_Total	Total distance, in yards, that all passes traveled.
PrgDist_Total	Total progressive distance, in yards, that all passes traveled towards the opponent's goal. Progressive passes are those that move the ball at least 10 yards closer to the opponent's goal or into the penalty area.
Cmp_Short	Number of completed short passes (5-15 yards).
Att_Short	Number of attempted short passes (5-15 yards).
Cmp_percent_Short	Percentage of short passes completed.
Cmp_Medium	Number of completed medium passes (15-30 yards).
Att_Medium	Number of attempted medium passes (15-30 yards).
Cmp_percent_Medium	Percentage of medium passes completed.
Cmp_Long	Number of completed long passes (more than 30 yards).
Att_Long	Number of attempted long passes (more than 30 yards).
Cmp_percent_Long	Percentage of long passes completed.
Ast	Number of assists credited to the player.
xAG	Expected Assisted Goals; the total xG of shots that result directly from a player's passes.
xA_Expected	Expected Assists; the likelihood that a given completed pass will become a goal assist.
A_minus_xAG_Expected	Difference between actual assists and expected assisted goals.
KP	Number of key passes; passes that directly lead to a shot.
Final_Third	Number of passes completed into the final third of the pitch.
PPA	Passes into the penalty area; number of passes completed into the 18-yard box.
CrsPA	Crosses into the penalty area; number of crosses completed into the 18-yard box.
PrgP	Progressive passes; number of completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area.

Table 9: Descriptions of Selected FBref Passing Variables

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
Mins_Per_90	Average minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
Att	Total number of passes attempted by the player.
Live_Pass	Number of live-ball passes attempted; passes made during active play, excluding dead-ball situations like free kicks or throw-ins.
Dead_Pass	Number of dead-ball passes attempted; passes made from set-piece situations such as free kicks, corners, or throw-ins.
FK_Pass	Number of passes made from free kicks.
TB_Pass	Number of through balls attempted; passes that penetrate the defensive line to find a teammate in an attacking position.
Sw_Pass	Number of switch passes; passes that change the side of attack by moving the ball from one flank to the opposite flank.
Crs_Pass	Number of crosses attempted; passes sent from wide areas into the penalty box.
TI_Pass	Number of throw-ins taken by the player.
CK_Pass	Number of corner kicks taken by the player.
In_Corner	Number of in-swinging corner kicks; corners that curve towards the goal.
Out_Corner	Number of out-swinging corner kicks; corners that curve away from the goal.
Str_Corner	Number of straight corner kicks; corners delivered without significant curve.
Cmp_Outcomes	Number of completed passes.
Off_Outcomes	Number of passes that resulted in the receiving player being offside.
Blocks_Outcomes	Number of passes that were blocked by an opponent.
Url	The specific URL link to the player's detailed statistics page on FBref.

Table 10: Descriptions of Selected FBref advanced Passing Variables

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
Mins_Per_90	Average minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
SCA_SCA	Total Shot-Creating Actions; the number of offensive actions (passes, dribbles, drawing fouls, etc.) that lead directly to a shot attempt.
SCA90_SCA	Shot-Creating Actions per 90 minutes; average number of shot-creating actions per 90 minutes played.
PassLive_SCA	Shot-Creating Actions from live-ball passes; number of shot-creating actions resulting from passes during active play.
PassDead_SCA	Shot-Creating Actions from dead-ball passes; number of shot-creating actions resulting from set-piece passes (corners, free kicks, throw-ins).
TO_SCA	Shot-Creating Actions from take-ons; number of shot-creating actions resulting from successful dribbles or take-ons.
Sh_SCA	Shot-Creating Actions from shots; number of shot-creating actions where a player's shot leads directly to another shot attempt (e.g., a saved shot leading to a rebound).
Fld_SCA	Shot-Creating Actions from fouls drawn; number of shot-creating actions resulting from fouls drawn by the player.
Def_SCA	Shot-Creating Actions from defensive actions; number of shot-creating actions resulting from defensive plays like tackles or interceptions.
GCA_GCA	Total Goal-Creating Actions; the number of offensive actions that lead directly to a goal.
GCA90_GCA	Goal-Creating Actions per 90 minutes; average number of goal-creating actions per 90 minutes played.

Table 11: Descriptions of Selected FBref Shot and Goal Creation Variables 1

Variable	Description
PassLive_GCA	Goal-Creating Actions from live-ball passes; number of goal-creating actions resulting from passes during active play.
PassDead_GCA	Goal-Creating Actions from dead-ball passes; number of goal-creating actions resulting from set-piece passes.
TO_GCA	Goal-Creating Actions from take-ons; number of goal-creating actions resulting from successful dribbles or take-ons.
Sh_GCA	Goal-Creating Actions from shots; number of goal-creating actions where a player's shot leads directly to a goal (e.g., a saved shot leading to a rebound goal).
Fld_GCA	Goal-Creating Actions from fouls drawn; number of goal-creating actions resulting from fouls drawn by the player.
Def_GCA	Goal-Creating Actions from defensive actions; number of goal-creating actions resulting from defensive plays like tackles or interceptions.

Table 12: Descriptions of Selected FBref Shot and Goal Creation Variables 2

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
Mins_Per_90	Average minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
Tkl_Tackles	Total number of tackles attempted by the player.
TklW_Tackles	Number of tackles won; instances where the player successfully dispossessed an opponent.
Def 3rd_Tackles	Number of tackles made in the defensive third of the pitch.
Mid 3rd_Tackles	Number of tackles made in the middle third of the pitch.
Att 3rd_Tackles	Number of tackles made in the attacking third of the pitch.
Tkl_Challenges	Total number of challenges (tackles plus duels) the player was involved in.
Att_Challenges	Number of challenges attempted by the player.
Tkl_percent_Challenges	Percentage of challenges that were successful.
Lost_Challenges	Number of challenges the player did not win.
Blocks_Blocks	Total number of blocks made by the player.
Sh_Blocks	Number of shots blocked by the player.
Pass_Blocks	Number of passes blocked by the player.
Int	Number of interceptions made by the player.
Tkl+Int	Combined total of tackles and interceptions made by the player.
Clr	Number of clearances made by the player; instances where the player clears the ball away from the defensive area.
Err	Number of errors committed by the player that led to an opponent's shot.
Url	The specific URL link to the player's detailed statistics page on FBref.

Table 13: Descriptions of Selected FBref Defensive Variables

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
Mins_Per_90	Average minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
Touches_Touches	Total number of touches by the player; each time the player makes contact with the ball.
Def_Pen_Touches	Number of touches in the defensive penalty area.
Def_3rd_Touches	Number of touches in the defensive third of the pitch.
Mid_3rd_Touches	Number of touches in the middle third of the pitch.
Att_3rd_Touches	Number of touches in the attacking third of the pitch.
Att_Pen_Touches	Number of touches in the attacking penalty area.
Live_Touches	Number of touches during live-ball situations; excludes set-pieces and dead-ball scenarios.
Att_Take	Number of take-on attempts; instances where the player tries to dribble past an opponent.
Succ_Take	Number of successful take-ons; instances where the player successfully dribbles past an opponent.
Succ_percent_Take	Percentage of successful take-ons.
Tkld_Take	Number of times the player was tackled during a take-on attempt.
Tkld_percent_Take	Percentage of take-on attempts where the player was tackled.
Carries_Carries	Total number of carries; instances where the player controls the ball with their feet and moves it.
TotDist_Carries	Total distance, in yards, that the player carried the ball.
PrgDist_Carries	Total progressive distance, in yards, that the player carried the ball toward the opponent's goal.

Table 14: Descriptions of Selected FBref Possession Variables 1

Variable	Description
PrgC_Carries	Number of progressive carries; carries that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any carry into the penalty area. Excludes carries which end in the defending 50% of the pitch.
Final_Third_Carries	Number of carries into the final third of the pitch.
CPA_Carries	Number of carries into the penalty area.
Mis_Carries	Number of times the player miscontrolled the ball during a carry, leading to a loss of possession.
Dis_Carries	Number of times the player was dispossessed by an opponent during a carry.
Rec_Receiving	Number of times the player was the target of a pass.
PrgR_Receiving	Number of times the player received a progressive pass; passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch.
Url	The specific URL link to the player's detailed statistics page on FBref.

Table 15: Descriptions of Selected FBref Possession Variables 2

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
MP_Playing.Time	Number of matches played by the player.
Min_Playing.Time	Total minutes played by the player.
Mn_per_MP_Playing.Time	Average minutes played per match.
Min_percent_Playing.Time	Percentage of total possible minutes played by the player.
Mins_Per_90_Playing.Time	Total minutes played divided by 90; represents the number of full matches played.
Starts_Starts	Number of matches started by the player.
Mn_per_Start_Starts	Average minutes played per start.
Compl_Starts	Number of matches where the player played the full duration.
Subs_Subs	Number of times the player was substituted into a match.
Mn_per_Sub_Subs	Average minutes played per substitution appearance.
unSub_Subs	Number of matches where the player was not substituted out.
PPM_Team.Success	Points per match earned by the team when the player was on the field.
onG_Team.Success	Goals scored by the team while the player was on the field.
onGA_Team.Success	Goals conceded by the team while the player was on the field.
plus/minus Goals	Goal differential (goals scored minus goals conceded) while the player was on the field. Real varibale name: plus_per__minus__Team.Success
plus/minus Goals per 90	Goal differential per 90 minutes while the player was on the field. Real variable name: plus_per__minus_90_Team.Success
On_minus_Off_Team.Success	Difference in goal differential per 90 minutes when the player is on the field versus off the field.
onxG_Team.Success..xG.	Expected goals (xG) for the team while the player was on the field.
onxGA_Team.Success..xG	Expected goals against (xGA) for the team while the player was on the field.
xG plus/minus	Expected goal differential (xG minus xGA) while the player was on the field. Real variable name: xG-plus_per__minus__Team.Success..xG
xG plus/minus per 90	Expected goal differential per 90 minutes while the player was on the field. Real variable name: xG-plus_per__minus_90_Team.Success..xG
On-Off xG	Difference in expected goal differential per 90 minutes when the player is on the field versus off the field. Real variable name: On_minus_Off_Team.Success..xG

Table 16: Descriptions of Selected FBref Playing Time and Team Success Variables

Variable	Description
Season_End_Year	The year in which the season concluded.
Squad	The name of the team or club the player is affiliated with.
Comp	The competition or league in which the player participated (e.g., Premier League, La Liga).
Player	The full name of the player.
Nation	The nationality of the player, often represented by a country code.
Pos	The primary position(s) the player occupies on the field (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the season's end.
Born	The birth year of the player.
Mins_Per_90	Average minutes played per 90 minutes; a measure of how many full matches' worth of minutes the player has played.
CrdY	Number of yellow cards received by the player.
CrdR	Number of red cards received by the player.
2CrdY	Number of instances where the player received a second yellow card in a match, leading to a red card.
Fls	Number of fouls committed by the player.
Fld	Number of times the player was fouled by an opponent.
Off	Number of times the player was caught offside.
Crs	Number of crosses attempted by the player.
Int	Number of interceptions made by the player; instances where the player intercepts an opponent's pass.
TklW	Number of tackles won; instances where the player successfully dispossesses an opponent.
PKwon	Number of penalties won by the player; instances where the player was fouled in the opponent's penalty area, resulting in a penalty kick.
PKcon	Number of penalties conceded by the player; instances where the player committed a foul in their own penalty area, resulting in a penalty kick for the opponent.
OG	Number of own goals scored by the player.
Recov	Number of ball recoveries by the player; instances where the player regains possession of a loose ball.
Won_Aerial	Number of aerial duels won by the player.
Lost_Aerial	Number of aerial duels lost by the player.
Won_percent_Aerial	Percentage of aerial duels won by the player.
Url	The specific URL link to the player's detailed statistics page on FBref.

Table 17: Descriptions of Selected FBref Miscellaneous Variables

Variable Name	Description
comp_name	Name of the competition or league in which the player participates.
region	Geographical region associated with the competition or league.
country	Country where the competition or league is based.
season_start_year	Starting year of the season for the given data.
squad	Name of the team or squad to which the player belongs during the season.
player_num	Jersey number assigned to the player.
player_name	Full name of the player.
player_position	Position on the field where the player primarily plays (e.g., Forward, Midfielder, Defender, Goalkeeper).
player_dob	Date of birth of the player.
player_age	Age of the player at the time of data collection.
player_nationality	Nationality or nationalities of the player.
current_club	Club for which the player is currently playing.
player_height_mtrs	Height of the player in meters.
player_foot	Dominant foot of the player (e.g., Left, Right, Both).
date_joined	Date when the player joined the current club.
joined_from	Previous club from which the player transferred.
contract_expiry	Date when the player's contract with the current club expires.
player_market_value_euro	Estimated market value of the player in euros.
player_url	URL to the player's profile on the Transfermarkt website.
MarketValue	Synonymous with player_market_value_euro; represents the player's market value in euros.
rownr	Row number or unique identifier for each record in the dataset.

Table 18: Description of Variables in the Transfermarkt Dataset

A.2 Additional PDP for all models

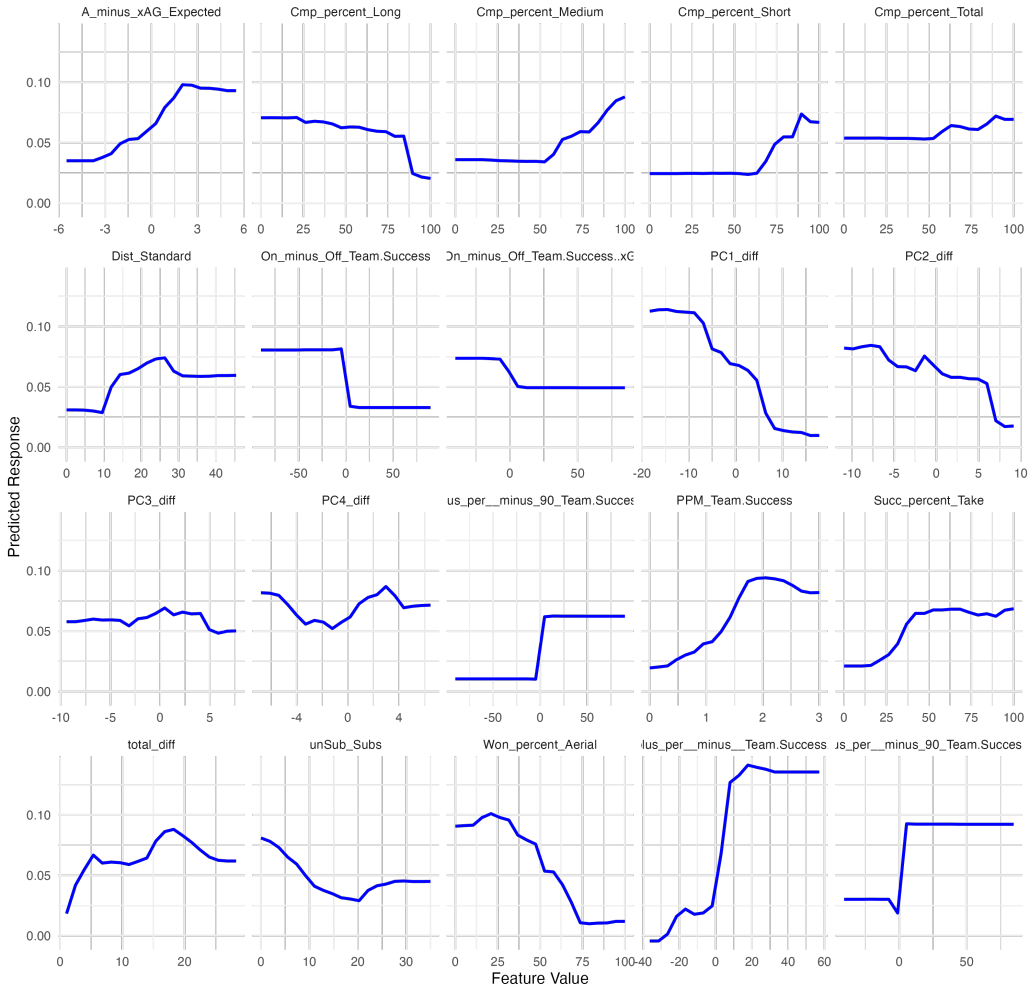

Figure 20: PDP for the top 20 variables for the the binary model for all positions (excluding categorical variables such as nation and the competition transferred to and from)

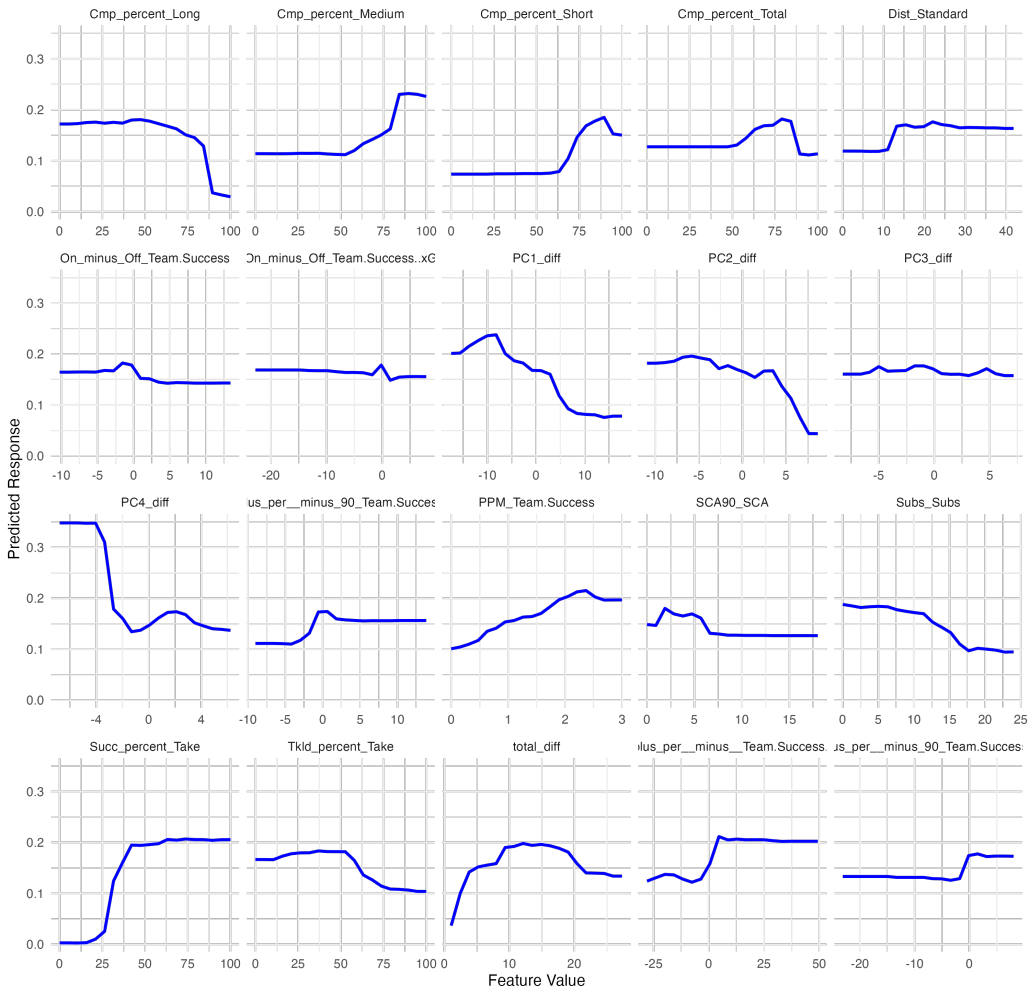

Figure 21: PDP for the top 20 variables for the the binary model for forwards (excluding the categorical variable: nation)

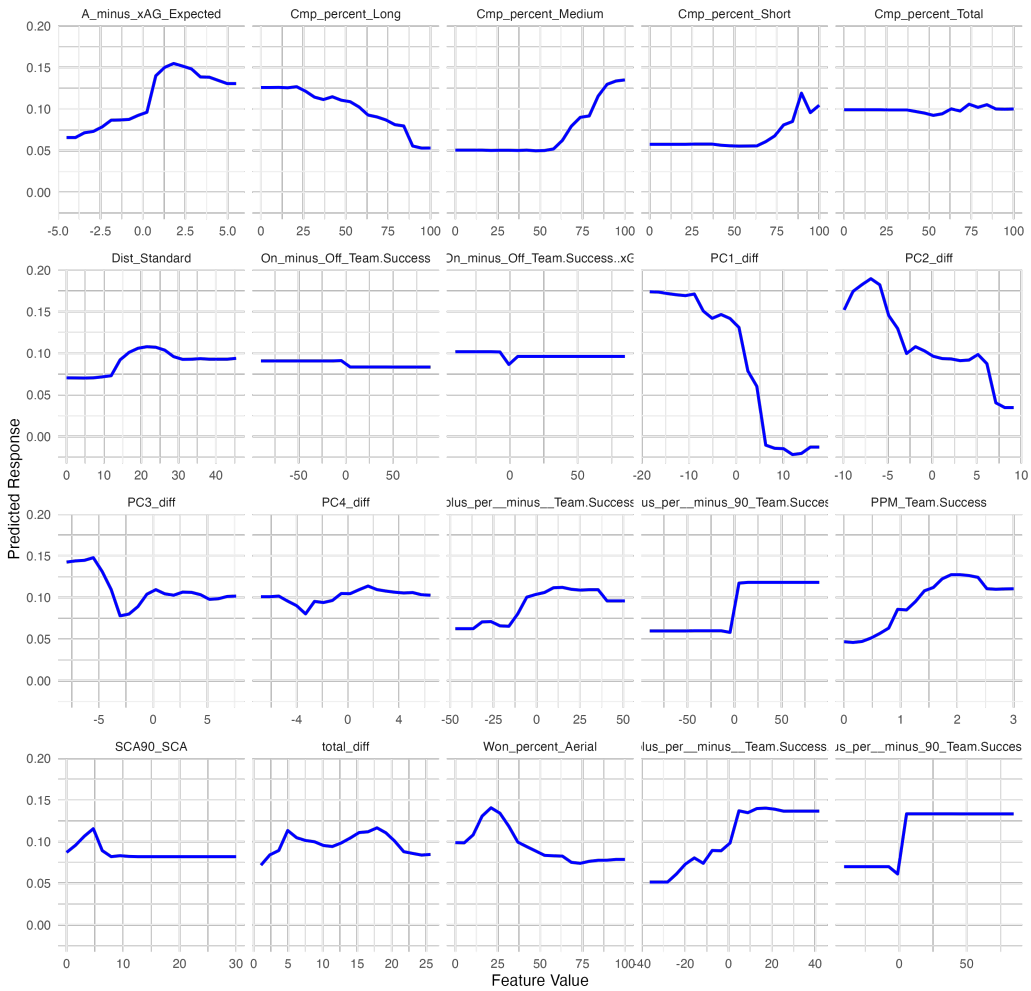

Figure 22: PDP for the top 20 variables for the the binary model for midfielders (excluding categorical variables such as nation and the competition transferred to and from)

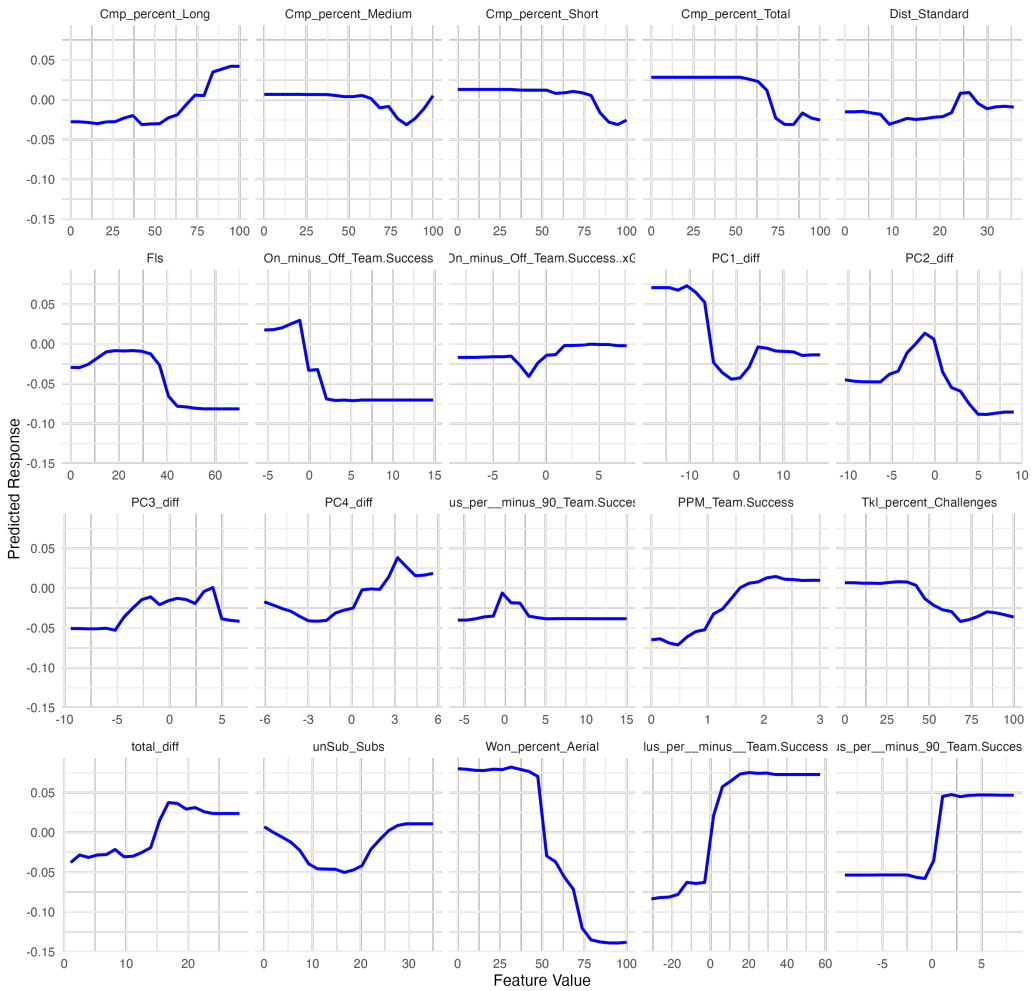

Figure 23: PDP for the top 20 variables for the the binary model for defenders (excluding categorical variables such as nation and the competition transferred to and from)

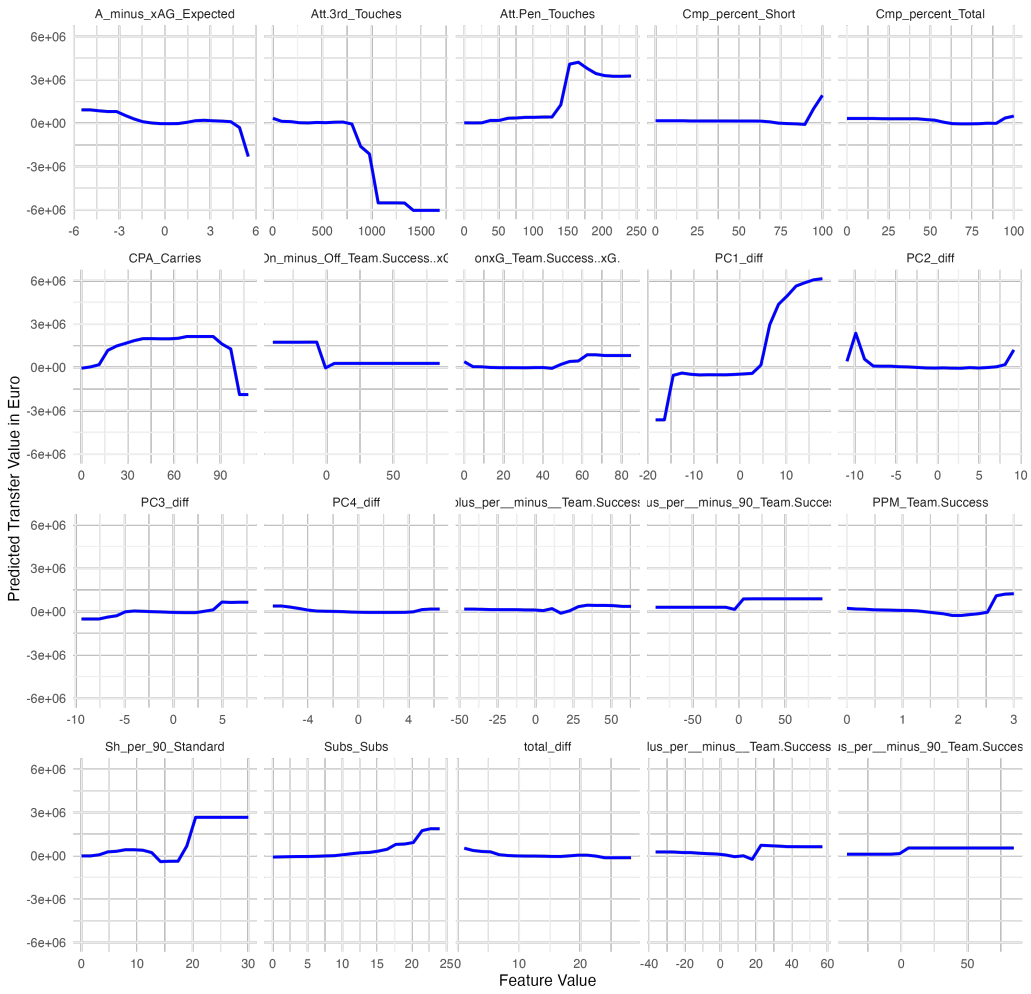

Figure 24: PDP for the top 20 variables for the the regression model for all positions (excluding the categorical variable: nation)

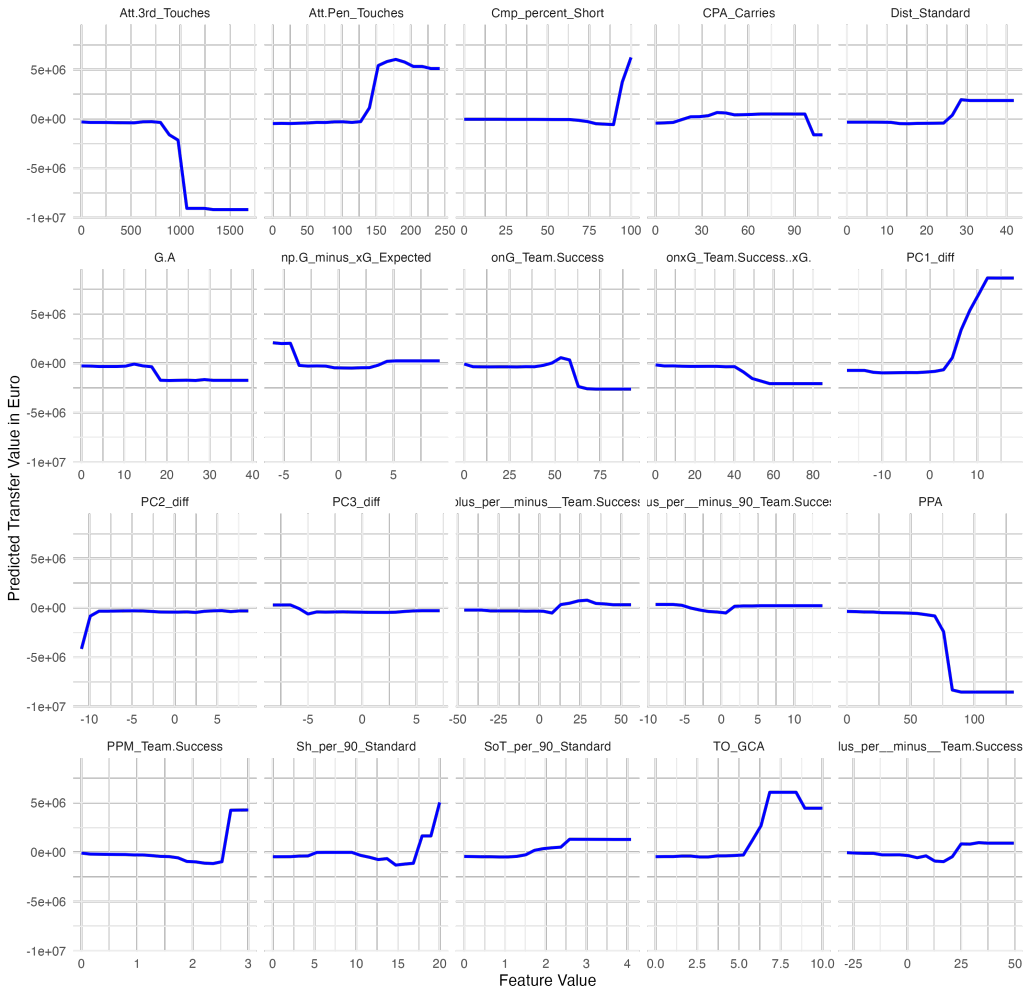

Figure 25: PDP for the top 20 variables for the the regression model for forwards

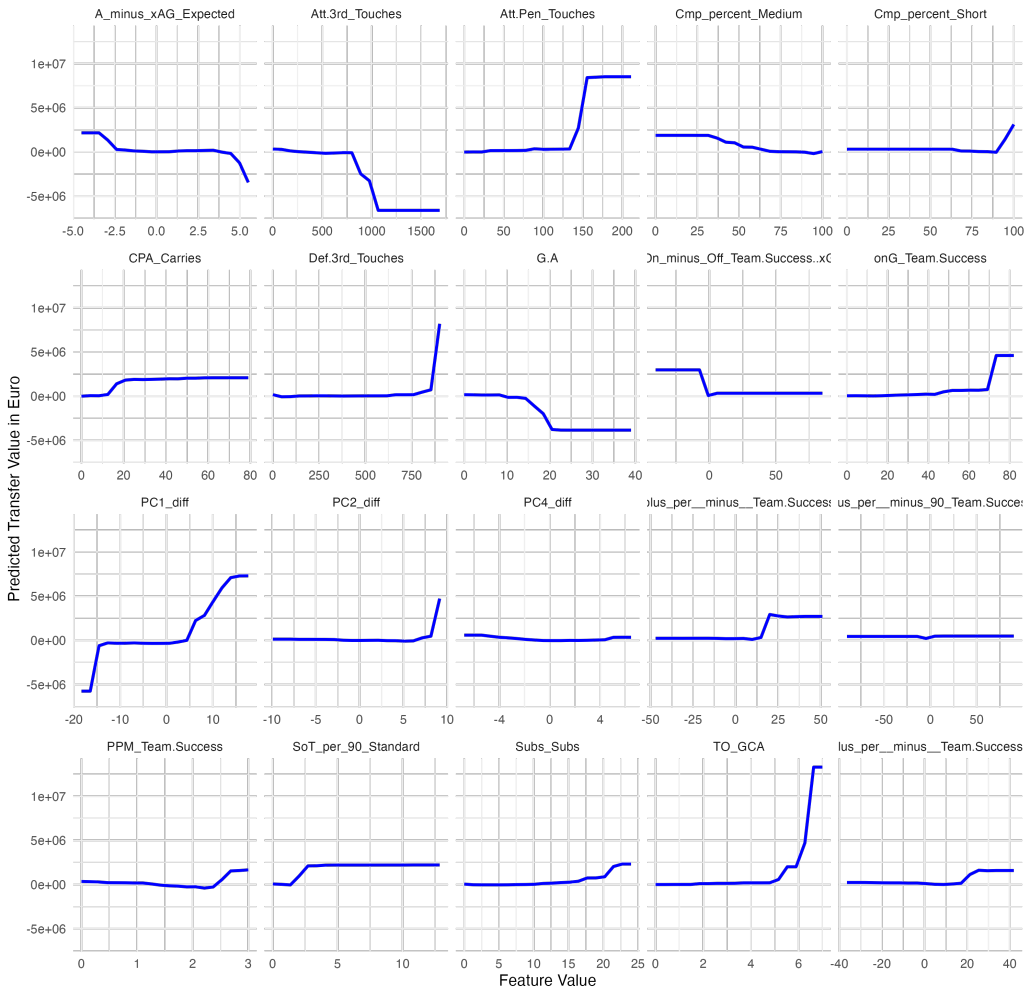

Figure 26: PDP for the top 20 variables for the the regression model for midfielders (excluding the categorical variable: nation)

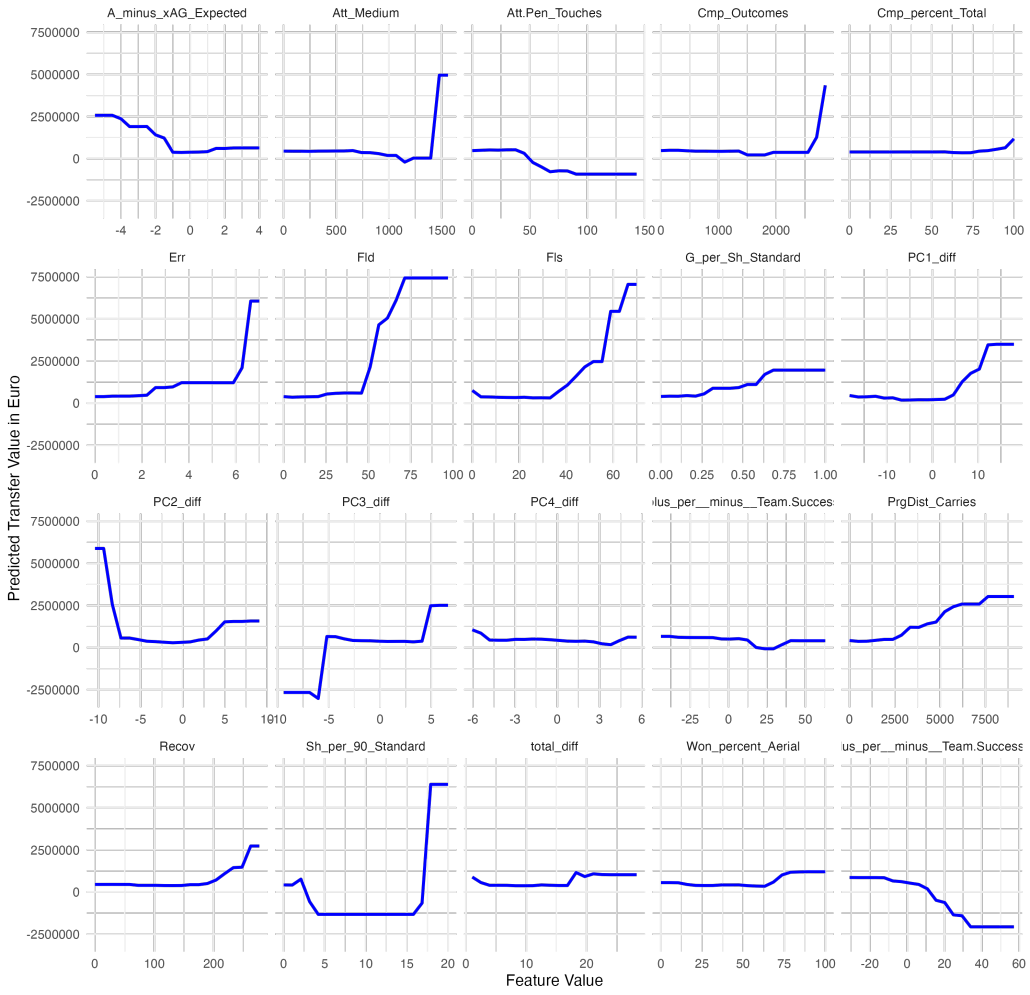

Figure 27: PDP for the top 20 variables for the the regression model for defenders

A.3 Variables used for PCA

Variable Name	Description
PrgC_Progression	Progressive Carries: Number of times a player carries the ball towards the opponent's goal at least 5 yards, excluding carries from the defensive 40% of the pitch.
PrgP_Progression	Progressive Passes: Completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, excluding passes from the defensive 40% of the pitch.
Att_Total	Passes Attempted: Total number of passes a player attempts.
TotDist_Total	Total Passing Distance: The sum distance, in yards, of all completed passes.
PrgDist_Total	Progressive Passing Distance: Total distance, in yards, that completed passes have traveled towards the opponent's goal.
Att_Medium	Medium Passes Attempted: Number of passes attempted that travel between 15 and 30 yards.
Att_Short	Short Passes Attempted: Number of passes attempted that travel less than 15 yards.
Att_Long	Long Passes Attempted: Number of passes attempted that travel more than 30 yards.

Table 19: Passing and Progression Statistics.

Variable Name	Description
Sh_Standard	Shots: Total number of shots taken by the player.
Dist_Standard	Average Shot Distance: The average distance, in yards, from which a player takes their shots.
KP	Key Passes: Passes that directly lead to a shot.
Final_Third	Passes into Final Third: Completed passes that enter the attacking third of the pitch.
PPA	Passes into Penalty Area: Completed passes into the 18-yard box.
CrsPA	Crosses into Penalty Area: Crosses that enter the 18-yard box.
Def.3rd_Tackles	Tackles in Defensive Third: Number of tackles made in the defensive third of the pitch.
Mid.3rd_Tackles	Tackles in Middle Third: Number of tackles made in the middle third of the pitch.
Att.3rd_Tackles	Tackles in Attacking Third: Number of tackles made in the attacking third of the pitch.
Att_Challenges	Attacking Challenges: Number of times a player challenges an opponent while they are on the attack.

Table 20: Shooting and Defensive Contributions.

Variable Name	Description
Def.Pen_Touches	Touches in Defensive Penalty Area: Number of times a player touches the ball within their own 18-yard box.
Def.3rd_Touches	Touches in Defensive Third: Number of times a player touches the ball in the defensive third.
Mid.3rd_Touches	Touches in Middle Third: Number of times a player touches the ball in the middle third.
Att.3rd_Touches	Touches in Attacking Third: Number of times a player touches the ball in the attacking third.
Att.Pen_Touches	Touches in Attacking Penalty Area: Number of times a player touches the ball within the opponent's 18-yard box.
Won_Aerial	Aerial Duels Won: Number of times a player wins an aerial duel.
Lost_Aerial	Aerial Duels Lost: Number of times a player loses an aerial duel.
Fls	Fouls Committed: Number of times a player commits a foul.
Fld	Fouls Drawn: Number of times a player is fouled by an opponent.
Off	Offsides: Number of times a player is caught offside.

Table 21: Touches, Aerial Duels, and Fouls.

B Electronic appendix

The R code for this thesis is available on:

https://github.com/justlampi/BT-SuccessfulFootballTransfer

References

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.
URL: *https://CRAN.R-project.org/package=gridExtra*

Bartosz, Ć., Giełczyk, A. and Choraś, M. (2021). Who will score? a machine learning approach to supporting football team building and transfers, *Entropy* **23**(1): 90.

Bhatnagar, P., Lokesh, G. H., Shreyas, J. and Flammini, F. (2024). Rating prediction of football players using machine learning, *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, pp. 121–126.

Bhavsar, K. A., Singla, J., Al-Otaibi, Y. D., Song, O.-Y., Zikria, Y. B. and Bashir, A. K. (2021). Medical diagnosis using machine learning: a statistical review, *Computers, Materials and Continua* **67**(1): 107–125.

Breiman, L. (2001). Random forests, *Machine learning* **45**: 5–32.

Bunker, R. and Thabtah, F. (2019). A machine learning framework for sport result prediction. applied computing and informatics, 15 (1), 27–33, *URL: http://www. sciencedirect. com/science/article/pii S* .

Cutler, A., Cutler, D. R. and Stevens, J. R. (2012). Random forests, *Ensemble machine learning: Methods and applications* pp. 157–175.

Deshpande, S. K. and Jensen, S. T. (2016). Estimating an nba player's impact on his team's chances of winning, *Journal of Quantitative Analysis in Sports* **12**(2): 51–72.

Dvorak, J., Graf-Baumann, T., Peterson, L. and Junge, A. (2000). Football, or soccer, as it is called in north america, is the most popular sport worldwide, *American journal of sports medicine* **28**(5 Suppl): S1–2.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of statistics* pp. 1189–1232.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles, *none* .

Gordon, A., Breiman, L., Friedman, J., Olshen, R. and Stone, C. J. (1984). Classification and regression trees., *Biometrics* **40**(3): 874.

Gough, C. (2024). Sports industry revenue worldwide in 2022, with a forecast for 2028, `https://www.statista.com/statistics/370560/worldwide-sports-market-revenue/`. [Online; accessed 17-March-2025].

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots, *The R Journal* **9**(1): 421–436.
URL: *https://doi.org/10.32614/RJ-2017-016*

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables with principal components, *J. Educ. Psy.* **24**: 498–520.

Hunt, T. (2020). *ModelMetrics: Rapid Calculation of Model Metrics*. R package version 1.2.2.2.
URL: *https://CRAN.R-project.org/package=ModelMetrics*

Inglis, A., Parnell, A. and Hurley, C. B. (2022). Visualizing variable importance and variable interaction effects in machine learning models, *Journal of Computational and Graphical Statistics* pp. 1–13.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments, *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **374**(2065): 20150202.

Jung, S. and Marron, J. S. (2009). Pca consistency in high dimension, low sample size context, *No Info* .

Kim, D. and You, K. (2023). Pca, svd, and centering of data, *arXiv preprint arXiv:2307.15213* .

Knoll, J. and Stübinger, J. (2020). Machine-learning-based statistical arbitrage football betting, *KI-Künstliche Intelligenz* **34**(1): 69–80.

Kong, L., Zhang, T., Zhou, C., Gomez, M.-A., Hu, Y. and Zhang, S. (2022). The evaluation of playing styles integrating with contextual variables in professional soccer, *Frontiers in psychology* **13**: 1002566.

Kuhn and Max (2008). Building predictive models in r using the caret package, *Journal of Statistical Software* **28**(5): 1–26.
URL: *https://www.jstatsoft.org/index.php/jss/article/view/v028i05*

Kursa, M. B. (2014). Robustness of random forest-based gene selection methods, *BMC bioinformatics* **15**: 1–8.

Lentz-Nielsen, N., Hart, B. and Samani, A. (2023). Prediction of movement in handball with the use of inertial measurement units and machine learning, *Sports Biomechanics* pp. 1–14.

Louppe, G. (2014). *Understanding random forests: From theory to practice*, PhD thesis, Universite de Liege (Belgium).

Ludkovski, M. (2023). Statistical machine learning for quantitative finance, *Annual Review of Statistics and Its Application* **10**(1): 271–295.

Luo, H. (2024). Improving nhl draft outcome predictions using scouting reports, *Journal of Quantitative Analysis in Sports* **20**(4): 331–349.

Ma, E. and Kabala, Z. J. (2024). Refereeing the sport of squash with a machine learning system, *Machine Learning and Knowledge Extraction* **6**(1): 506–553.

Maćkiewicz, A. and Ratajczak, W. (1993). Principal components analysis (pca), *Computers & Geosciences* **19**(3): 303–342.

Mead, J., O'Hare, A. and McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value, *Plos one* **18**(4): e0282295.

Molnar, C. (2020). *Interpretable machine learning*, Lulu. com.

Newman, J., Sumsion, A., Torrie, S. and Lee, D.-J. (2023). Automated pre-play analysis of american football formations using deep learning, *Electronics* **12**(3): 726.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2**(11): 559–572.

Pengyu, W. and Wanna, G. (2021). Image detection and basketball training performance simulation based on improved machine learning, *Journal of Intelligent & Fuzzy Systems* **40**(2): 2493–2504.

Plakias, S., Kokkotis, C., Moustakidis, S., Tsatalas, T., Papalexi, M., Kasioura, C., Giakas, G. and Tsaopoulos, D. (2023). Identifying playing styles of european soccer teams during the key moments of the game, *Journal of Physical Education and Sport* **23**(4): 878–890.

Puerzer, R. J. (2002). From scientific baseball to sabermetrics: Professional baseball as a reflection of engineering and management in society, *NINE: A Journal of Baseball History and Culture* **11**(1): 34–48.

Rai, K., Wang, Y., O'Connell, R. W., Patel, A. B. and Bashor, C. J. (2024). Using machine learning to enhance and accelerate synthetic biology, *Current Opinion in Biomedical Engineering* p. 100553.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves, *BMC Bioinformatics* **12**: 77.

Schulte, O., Khademi, M., Gholami, S., Zhao, Z., Javan, M. and Desaulniers, P. (2017). A markov game model for valuing actions, locations, and team performance in ice hockey, *Data Mining and Knowledge Discovery* **31**: 1735–1757.

Shen, B., Shalaginov, M. Y. and Zeng, T. H. (2023). Injury risk prediction in soccer using machine learning, *2023 International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 2103–2106.

Stathead.com (2025). Stathead - sports statistics database. Accessed: 2025-03-21.
URL: *https://www.stathead.com/*

Strobl, C., Boulesteix, A.-L. and Augustin, T. (2007). Unbiased split selection for classification trees based on the gini index, *Computational Statistics & Data Analysis* **52**(1): 483–501.

Transfermarkt.com (2025). Transfermarkt - football transfers stats. Accessed: 2025-03-21.
URL: *https://www.transfermarkt.com/*

Wei, M., Zhong, Y., Zhou, Y., Gui, H., Yu, S., Yu, T., Guan, Y. and Wang, G. (2023). Research progress of sports injury prediction model based on machine learning, *International Symposium on Computer Science in Sport*, Springer, pp. 23–41.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. and Yutani, H. (2019). Welcome to the tidyverse, *Journal of Open Source Software* **4**(43): 1686.

Wickham, H. and Henry, L. (2025). *purrr: Functional Programming Tools*. R package version 1.0.4.
URL: *https://CRAN.R-project.org/package=purrr*

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R, *Journal of Statistical Software* **77**(1): 1–17.

Declaration of authorship

I hereby declare that the Thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This thesis was not previously presented to another examination board and has not been published. ChatGPT-4o, ChatGPT Internet Search and GitHub Copilot were used to aid this work. This served as assistance to my own ideas, providing no new ideas or creation of new text. Improvement of language and grammar was however done with Chat-GPT-4o, keeping the meaning and context of the text. ChatGPT-4o and Github Copilot were used to aid in writing code, while Github Copilot is integrated into my RStudio environment. ChatGPT-4o was also used for LaTeX code.

Munich, March, 24th 2025

Justin Lampman