TILBURG UNIVERSITY

NEVER CHANGE A WINNING TEAM, OR SHOULD YOU?

ASSESSING THE IMPACT OF COACH INTERVENTIONS IN PLAYER LINEUPS IN SPORTS USING MACHINE LEARNING MODELS

CHRISTOFFEL A.S. RÜGER

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

840567

COMMITTEE

prof. dr. Barbara Deleersnyder
dr. Drew Hendrickson

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 20, 2023

WORDS

8,765

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my thesis supervisor prof. dr. Barbara Deleersnyder, for her inestimable guidance, support, and encouragement throughout my research journey. I would also like to thank my teachers, family, and friends who all helped me during my education.

NEVER CHANGE A WINNING TEAM,OR SHOULD YOU?

ASSESSING THE IMPACT OF COACH INTERVENTIONS IN PLAYER LINEUPS IN SPORTS USING MACHINE LEARNINGMODELS

Christoffel A.S. Rüger

Abstract

In this thesis, we proposed a model splitting soccer matches per minute in order to examine if both predetermined match features, as well as dynamic in-game features, predict the final number of goals the home and away team will score. We used the data of the four biggest national soccer leagues (Bundesliga, La Liga, Serie A, and Premier League) from the seasons 18/19 till 21/22 to train and test the predictions of a multilayer perceptron and compared the performance with multiple linear regression and random forest. Furthermore, we examined the impact a coach can have on the final number of goals by making interventions in the player lineup during the game. The results demonstrate that overall, multilayer perceptron outperformed multiple linear regression (MSE -0.053) and random forest (MSE -0.014) with an increased accuracy of 1.42% and 0.41% respectively. Including the number of substitutions used and the change in strategy score significantly reduced the mean squared error by 0.22. Furthermore, increasing the number of substitutions used during the game significantly decreased the final number of goals for both the home and away team. Contrary to expectations, changing to a more attacking lineup by the home team decreases the final number of goals of both the home and away team, and changing to a more defensive lineup by the home team increases the final number of goals of both the home and away team.

Data Source/Code/Ethics Statement

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data or code. The code used in this thesis is publicly available via https://github.com/casruger/Data-Science-and-Sociaty-Thesis. All figures and images are made by and belong to, the author hence no licensing was necessary.

1. Introduction

Recently, live in-game prediction models have become more popular in predicting sports outcomes. These models use both predetermined match factors as well as dynamic in-game factors to predict the match winner (Thabtah, Zhang, & Abdelhamid, 2019).

Coaches try to increase their team's winning chances as much as possible before, but also during matches. They can do so by giving instructions to the players during a time-out, but a more direct way to influence the game in a team sport is by substituting certain players and changing the team's strategy that way.

The more recent advances in prediction models where researchers include dynamic in-game features in predicting sports outcomes offer the opportunity to explore the effects of these strategic coach changes on match outcomes. The interventions of coaches on player lineups (changing a more attacking player for a more defensive one and the other way around) are examples of such dynamic in-game features which have not yet been thoroughly researched.

Especially in soccer, models which take into account both predetermined as well as dynamic in-game features are not as widely used due to the high number of even (tie) match outcomes (Robberechts, Haaren & Davis, 2021). Therefore, this thesis focuses on the number of goals the home and the away soccer team will score rather than predicting who will be the match-winner. Using the number of goals as a dependent variable has the advantage of finding relevant features more quickly, as the number of goals contain more information about a team's strength than simply the category of win, tie, or loss (Ganguly & Frank, 2018). For example, a 2-1 victory is not as convincing as a 10-1 victory. In addition, matches with more goals are generally considered more exciting (Brocas & Carrillo, 2004). Therefore, increasing the number of goals could enhance the popularity of watching the sport.

The aim of this thesis is to find a new way of using an artificial neural network that takes both predetermined features as well as dynamic in-game features to predict the final number of home and away goals separately. In addition, this thesis aims to research the influence of coach interventions in player lineups on the number of goals for the home and away team.

1.1. Problem statement & research questions

Against this background, the research questions this thesis seeks to answer are:

1. To what extent can the final number of goals (for the home and away team) be predicted using a combination of predetermined match factors and dynamic in-game features during the match?
    a. How accurately can an artificial neural network predict the final number of goals per match (for the home and away team) using a combination of predetermined match factors and dynamic in-game features during the match, compared to machine learning models such as linear regression and random forest?
2. To what extent do coach interventions in player lineups during soccer matches influence the final number of goals for the home and away team?
    a. To what extent does changing to a more defensive player lineup affect the number of predicted goals for the home and away team in a soccer match?

b. To what extent does changing to a more attacking player lineup affect the number of predicted goals for the home and away team in a soccer match?

1.2. Academic and managerial relevance

1.2.1. Academic relevance

For soccer there have been multiple studies predicting the final match winner based on predetermined factors as well as dynamic in-game features retrieved from the first half (Razali et al., 2017; Capobianco et al., 2019). Robberechts, Haaren & Davis (2021) also predicted the match winner based on predetermined factors as well as dynamic in-game features, but rather than using only one time interval (after the first half) they split each match into 100-time intervals. This thesis will add to the existing literature a new way of predicting the number of goals the home and away team will score using both dynamic in-game features in addition to predetermined match features rather than predicting the match winner as the existing literature did.

In addition, this research will apply an artificial neural network which in previous soccer research has only been applied to datasets and matches containing only predetermined match factors (Igiri & Nwachukwu, 2014; Rudrapal, Srivastava & Singh, 2020).

Lastly, this research will look at the influence of coach interventions in player lineups on the final number of goals for the home and away team. Previous research has looked, for instance, at the reduction in fatiguing effects by using substitutions (Liu et al., 2019). However, the effects of tactical substitutions, for example changing an attacker for a defender, have yet to be researched.

1.2.2. Managerial relevance

The outcome of this research has relevance to several stakeholders. Firstly, soccer clubs and their coaches could benefit by improving their lineup during matches based on real-time predictions made by artificial neural networks. The findings can be directly implemented by teams to make better decisions based on empirical evidence on the effect of specific coach interventions.

Secondly, competition hosts such as the Federation Internationale de Football Association (FIFA) could benefit from this research by making the matches more exciting by setting or adjusting the rules of the game. In Formula 1, for example, you must use at least two types of tires. This rule has been implemented to make the sport more exciting for the viewer. If certain coach interventions increase the number of goals and thus raise the perceived excitement of a match, competitions could implement a rule encouraging teams and coaches to make such interventions.

Lastly, this thesis has implications for viewers of team sports and society as a whole. If the number of goals per match could increase due to better coach interventions and the implementation of new rules, this would enhance the viewing experience since the matches are more exciting. This would in turn increase the number of viewers and watch time. Consuming sports media has been linked to improved well-being, less loneliness, and higher self-esteem (Kim et al., 2017; Council of Europe, 2022). Therefore, increasing the number of goals per match and the excitement of matches could amplify these benefits for current viewers by making them watch more matches, and also attract new viewers, providing them these benefits as well.

1.3. Research strategy

In order to address the research questions, a comprehensive literature review will be conducted and summarized in the next chapter. This review will include related works and current model benchmarks. In chapter three, the research methods will be chosen and explained in detail. This includes the use of multiple linear regression, random forest, and multilayer perceptron and the criteria used to evaluate and validate the models. Data will be collected from two sources, Sportmonks and Clubelo.com, and will be combined into a single dataset for training, validation, and testing of the models. The details of data collection, preparation, hardware/software used, and exploratory analysis of the data will be provided in chapter four. The assumptions of the models will be tested to ensure the results can be interpreted correctly. If the assumptions are met, the results of the different models will be evaluated and compared using the criteria outlined in chapter three. An error analysis will be conducted to assess the robustness of the models and to check for any disparate bias. As soccer data is time-ordered, it is expected that the model prediction performance will increase during the season and during the game. Whether this or other unintentional discriminations are made by the model will be verified. More details on the error analysis will be provided in chapter five. Chapter five also contains different feature selection tests, to see whether the model improves by selecting different features and/or modifying them. From the results, conclusions will be drawn and linked back to the research questions. Lastly, a discussion will be provided outlining the limitations and directions for future research.

2. Background literature

2.1. Predicting sports outcomes using predetermined match factors

Prediction models for both individual and team sports have outperformed human experts in all kinds of sports ranging from predicting the distance an athlete will throw a javelin (Edelmann-Nusser, Hofmann & Henneberg, 2002) to predicting who will win a basketball match (Huang Lin, 2020). Joseph, Fenton, and Neil (2006) researched the use of multiple Bayesian analyses to predict a single team's soccer match outcome. They found that for predicting a single team's match outcome, the Bayesian network with an accuracy of 59% outperforms MC4 decision trees, and K-nearest neighbor models.

Other researchers used different models. Igiri and Nwachukwu (2014) built a logistic regression and applied a neural network to predict the match winner and produced a prediction accuracy of 95%. The extremely high accuracy of Igiri and Nwachukwa was questioned by Prasetio and Harlili (2016). They used the same models but on a bigger sample, and achieved an accuracy of only 69%.

More recently, Stübinger, Mangold, and Knoll (2019) used multiple machine learning models on over 45.000 matches from the five biggest soccer leagues to predict match outcomes. They found that random forest, with an accuracy of 81%, outperformed other machine learning models such as gradient boosting, support vector machine, and linear regression. This is in line with the findings of Fahey-Gilmour et al. (2019), who also found that the random forest outperformed other machine learning models such as logistic regression, single hidden layer neural network, support vector machine, and gradient boosting in predicting soccer match outcomes

In addition to machine learning models, Rudrapal, Srivastava, and Singh (2020) trained a deep learning multilayer perceptron and found that the multilayer perceptron with an accuracy of 73% outperformed random forest, support vector machine, and Gaussian Naïve Bayes. This implies that deep learning might be better at predicting match outcomes using predetermined match features than more traditional machine learning models. Table 1 provides a summary of the studies predicting match outcomes in soccer games.

Table 1. Overview of the literature predicting soccer outcomes using predetermined match factors

Source	**Models**	**Accuracy**	**Dependent variable**
Joseph, Fenton & Neil, 2006	**Bayesian network,** MC4 decision trees, K-nearest neighbor	59%	Probability of the home team to win, or draw
Igiri & Nwachukwu, 2014	**Logistic regression**, Artificial neural network	95%	Probability of the home team to win
Prasetio & Harlili, 2016	**Logistic regression**	69%	Probability of the home team to win
Razali, Mustapha, Ahmad & Ruhaya, 2017	**Bayesian network**	75%	Probability of the home team to win, or draw
Stübinger, Mangold &Knoll, 2019	**Random forest,** Gradient boosting, Support vector machine, Linear regression	81%	Probability of the home team to win
Fahey-Gilmour, Dawson, Peeling, Heasman & Rogalski, 2019	**Random forest,** Logistic regression, Single layer perceptron, Support vector machine Gradient boosting	72%	Probability of the home team to win, or draw
Rudrapal, Srivastava & Singh, 2020	**Multilayer perceptron,** Random forest, Support vector machine, Gaussian Naïve Bayes	73%	Probability of the home team to win, or draw

2.2. Prediction models using both predetermined as well as dynamic in-game features

In recent years, the popularity of models predicting sports outcomes by using both predetermined features as well as dynamic in-game features has increased. An overview of this literature is offered in Table 2. Bailey and Clarke (2006) build a model predicting the match outcome of one-day international cricket matches while the game is in progress. Using the information gained in the first inning, they achieved an accuracy of 71% on predicting the final match winner using multiple linear regression. Similarly, Kumar and Roy (2018) researched whether they could predict the total number of points scored in a running inning. They used the combination of venue variables and dynamic in-game features and found that the multilayer perceptron slightly outperformed the multiple linear regression. More recently, Bhawkar and Patel (2019) made a model predicting the match winner for the Indian Premier cricket league. They did so for four-time series, namely before the match, before the 5th inning, 10th inning, and 15th inning. They also found that their multilayer perceptron outperformed logistic regression, Gaussian Naïve Bayes, and linear regression.

Shirley (2007) was the first to propose a model splitting each game not in four quarters, but in within-game events. Štrumbelj and Vračar (2012) built upon this model by including team-specific variables to account for the individual team's strength and used this to predict the match winner of basketball matches. Their logistic regression model achieved an accuracy of 69%. Merritt and Clauset

(2014) extended the model of Štrumbelj and Vračar by including the number of seconds left for each match and found that the prediction accuracy increased to 80% after only 40 events. Kayhan and Witkins (2018) split each game into 2880 snapshots, each representing 1 second of the match. They found that the model containing only the goal difference between home and away was the best-performing model with an average Brier score of 0.170. Kayhan and Witkins (2019) continued their work by comparing the nearest neighbor snapshot approach to a linear regression model and a long short-term memory network and found that the long short-term memory network performed worse than the other two, but the snapshot approach and general linear model did not significantly differ in their predictions.

In soccer, the literature on prediction models taking both predetermined as well as dynamic in-game features is more scarce than in other sports like cricket and basketball. Razali et al. (2017) build upon the Bayesian soccer analysis of Joseph et al. (2006) by including the values of the following dynamic in-game features up until half-time per team; shots, shots on target, corners, fouls, yellow cards, red cards, and goals. They increased the accuracy from 59% to 75% using the same Bayesian Network. Capobianco et al. (2019) proposed a model using different variables, namely ball possession, distance covered, the center of gravity of the team, and the number of ball recoveries from the first half of a soccer match to predict the final match outcome. They found that the random forest with an average precision of 0.843 outperformed the random tree and multilayer perceptron. However, they used only 1 season, which is a relatively small sample size. Recently, research by Robberechts, Haaren, and Davis (2021) build a linear regression and a random forest model to predict soccer match outcomes. They are the first and only researchers who predicted the final match outcome at multiple stages during the game. Rather than predicting the final match outcome during half-time, they split each match into 100-time intervals and predicted the final match outcome for each time interval resulting in 100 predictions. They found that the random forest had a higher accuracy compared to both linear and multiple linear regression models. They did not include any deep learning models even though they had over 6,000 matches in their dataset, which according to Xu et al. (2021), could suit deep learning models better. Furthermore, they did not include any features related to coach interventions in player lineups which could improve the accuracy of the models. An overview of this literature is offered in Table 2.

Table 2. Overview of the literature predicting sports outcomes using both predetermined and dynamic in-game features

Source	**Models**	**Accuracy**	**Sport**	**In-game features time interval**	**Dependent variable**
Bailey & Clarke, 2006	Multiple linear regression	71%	Cricket	First inning	Probability of home win, home loss, or draw based on the predicted number of points (home and away)
Kumar & Roy, 2018	**Multilayer perceptron,** Multiple linear regression	MSE: 41	Cricket	First inning	Total number of points of both the home and away team
Bhawkar & Patel, 2019	**Multilayer perceptron,** Logistic regression, Gaussian Naïve Bayes, Linear regression	76%	Cricket	5th inning, 10th inning & 15th inning	Probability of home win or home loss or draw based on the predicted number of points (home and away)
Štrumbelj & Vračar, 2012	**Logistic regression**	69%	Basketball	Per in-game event	Probability of home win or home loss
Merritt & Clauset, 2014	**Logistic regression**	80%	Basketball	Per in-game event, including each minute	Probability of home win or home loss
Kayhan & Witkins, 2018	**Nearest neighbor snapshot**	Brier: 0.170	Basketball	Each second	Probability of home win or home loss
Kayhan & Witkins, 2019	**Multiple linear regression,** **Nearest neighbor snapshot,** Long short-term memory network	MAE: 8	Basketball	Each second	Total number of points (home and away team)
Razali, Mustapha, Ahmad & Ruhaya, 2017	**Bayesian network**	75%	Soccer	Half time	Probability of home win, home loss, or draw
Capobianco, Di Giacomo, Mercaldo, Nardone & Santone, 2019	**Random forest,** Random tree, Multilayer perceptron	Precision: 0.857 Recall: 0.750	Soccer	Half time	Probability of home win or home loss
Robberechts, Haaren & Davis, 2021	**Random forest,** Linear regression, Multiple linear regression	74%	Soccer	100-time intervals per match	Probability of home win, home loss, or draw based on the predicted number of goals (home and away)

2.3. The Role of coach interventions in player lineups

Most research has looked at the change in physiological parameters due to player substitutions. Liu et al. (2019) found that the substituted players covered more ground and spent more time jogging rather than walking per minute compared to players who played the entire match. This is in line with previous findings that substituting players reduces the effects of fatigue across the team (Bradley et al., 2014; Hills et al., 2018). What effects substituting a player has on the number of goals per match or the winning chances has yet to be researched.

In addition to preventing fatigue effects, coaches can substitute players in order to change the strategy of the team. Miguel et al. (2020) investigated the effect of player substitutions on tactical behavior in high-performance soccer. They analyzed the 2016-2017 Bundesliga and found that when a team is behind, they are far more likely to substitute a defensive player for a more attacking player (48.1%) rather than substituting for a player with the same playing style (19.8%). This is contrary to earlier beliefs of ‘never change a winning team’ (A. Ramsey, June 26, 2009). Miguel et al. also found that changing to a more defensive playing style decreased the team's centroid, team length, and space control. Changing to a more offensive strategy had the opposite effect. Rein et al. (2017) found that soccer teams with greater space control have been associated with a higher probability of winning the game. This implies that changing to a more offensive playing style increases the team's winning chances. However, whether this is the case or whether any of these coach interventions have an influence on the number of goals has yet to be examined empirically since they have not been included in models predicting match outcomes using both predetermined match factors as well as in-game features.

3. Methodology

This chapter starts with an overview of the variables that will be included in the prediction models. Furthermore, an explanation is given for the chosen models. In addition, the model evaluation criteria are explained to select the best-performing prediction model. Lastly, the evaluation and validation procedure of the models are described.

3.1. Dependent variable: Final number of goals for the home and away team

Following the literature, the most common dependent variable for predicting match outcome is by grouping the match in either home win, home loss, or draw (Kayhan & Witkins, 2019; Capobianco, et al., 2019). The downside of this classification is that it is harder to find significant independent variables for variables with small changes. This is due to the information loss in classifying the game outcomes only based on 3 outcomes (Ganguly & Frank, 2018). Predicting the final number of goals of the home and away team results in more variation in the dependent variable and hence, more powerful prediction models compared to models predicting the win probability. Two matches with final scores of 10-1 and 2-1 would both be classified the same as home wins with the match winner as dependent variable and thus, both update the model estimations equally. However, these games should be treated differently since the first game was more convincing. In addition, this study is more interested in the final number of goals rather than predicting the winner. For these reasons the dependent variable on which the model will be trained and tested is the final number of goals of both the home and away team separately.

3.2. Independent variables

To predict the final number of goals, multiple independent variables are included in the models which have been shown in other studies to have an effect on the dependent variables. An overview of all variables can be seen in Table 3.

3.2.1. Predetermined match features

For each match, the league and the season are coded as categorical variables. In addition, the date of the match is collected. This date is used to split the train and test data but is not included in the models as an independent variable. In line with previous research, the quality of both teams is captured and included in the models using the Elo rating system (Robberechts, Haaren & Davis, 2021) since this has been found to be a significant predictor of match outcomes and is a proven measurement of team's strength (Hvattum &Arntzen, 2010). Secondly, in team sports like soccer, the home team has been found to win significantly more games than the away team (Pollard, 2008). This is accounted for by having two separate models predicting the final number of home and away goals.

3.2.2. Dynamic in-game features

Every match is split into 90 rows all counting for 1 minute of the game. For each minute, the dynamic in-game features are updated according to the match events. The number of red and yellow cards have been found to have an influence on the final match outcome (Robberechts, Haaren & Davis, 2021) and is therefore included as an independent variable. The same holds for the current number of goals since this again has a significant influence on the final match outcome. In addition, two dynamic coach interventions in player lineup features are included.

First, the number of player substitutions during the game is added. Hills et al. (2018) found that substituting a player reduces the fatiguing effects and could therefore influence the final number of goals per team. The second feature is the change in team strategy since this is expected to have an effect on the team's tactical behavior (Rein et al., 2017). The teams' strategy score is set at 0 at the start of the game and increases if a more attacking player is substituted in and decreases if a more defensive player is substituted in during the game. More information on the data gathering and the calculation of team strategy can be found in chapter 4.

Table 3. Overview of variables and what they measure.

Variable	**Measured**
Dependent variables	*(value fixed for each game)*
Final number of home goals	Number of goals the home team scored at the end of the match
Final number of away goals	Number of goals the away team scored at the end of the match
Independent variables	
Predetermined match features	*(value fixed for each game)*
League	The league as categorical variable
Season	The season as categorical variable
Change in Elo	The difference in Elo rating (home – away)
Dynamic in-game features	*(measured every minute in the game)*
Minute	The cumulative minutes already played in the match
Home goals	The cumulative number of goals the home team has scored
Away goals	The cumulative number of goals the away team has scored
Home yellow cards	The cumulative number of yellow cards the home team has received
Away yellow cards	The current number of yellow cards the away team has received
Home red cards	The current number of red cards the home team has received
Away red cards	The current number of red cards the away team has received
Coach interventions features	*(measured every minute in the game)*
Number of home substitutions	The cumulative number of substitutions used by the home team
Number of away substitutions	The cumulative number of substitutions used by the away team
Home strategy	The current strategy score for the home team, starts at 0 and increases when a more attacking player is substituted in, decreasing when a more defensive player is substituted in.
Away strategy	The current strategy score for the away team, starts at 0 and increases when a more attacking player is substituted in, decreasing when a more defensive player is substituted in.

3.3. Overview prediction models and model selection

Tables 1 and 2 in chapter 2 show that the most commonly used prediction models in sports games are logistic regression and random forest. Logistic regression is an easy-to-compute and interpretable linear classification model, but it can only predict a binary outcome (win or not). In this research, the predicted variable (number of goals) is not binary, but it can take positive values of 0 and higher, making the use of logistic regression inadequate. Linear regression is built upon the same principles as logistic regression, but it infers the relationship between a continuous dependent variables and the independent variables that can be both continuous or categorical. In this thesis, multiple independent variables are expected to have an effect on the dependent variable. Therefore, the multiple linear regression will be estimated. In addition, multiple linear regression is ideal for esteblasing a baseline (Abramovich & Ritov, 2013), and has been proven to predict match outcomes very well (Bailey & Clarke, 2006; Kayhan & Witkins, 2019).

The second most often used model to predict sport outcomes is random forest. Contrary to linear regression, random forests can find non-linear relationships. Random forest has outperformed other machine learning models such as gradient boosting, support vector machine, and Gaussian Naïve Bayes in predicting soccer match outcomes (Bhawkar & Patel, 2019, Fahey-Gilmour et al., 2019). Since random forest is currently the best-performing model predicting soccer match outcomes using both predetermined as well as in-game features with multiple time intervals, this model is estimated as well.

Multilayer perceptron has been shown to be an effective alternative to more traditional statistical techniques (Schalkoff, 1992). Unlike other statistical techniques, multilayer perceptron makes no prior assumptions concerning the data distribution. It can model highly non-linear functions and can be trained to accurately generalize when presented with new, unseen data (Gardner & Dorling, 1998). In addition, in other sports, multilayer perceptron has achieved the highest accuracy in predicting the final score for a team (Bhawkar & Patel, 2019; Kayhan & Witkins, 2019). Therefore, the last model that is retained in this study is the multilayer perceptron.

3.4. Multiple linear regression

Multiple regression analysis is a statistical learning technique that infers the relationship between the dependent variable, and one or more independent variables (Cook, 2001). In this study, multiple fixed and dynamic independent variables are hypothesized to influence the final number of goals per team. Linear regression is one of the oldest and most often implemented approaches in statistics. Linear models are easy to handle mathematically and provide adequate and interpretable estimations of the relationships between the independent and dependent variables. These traits make them ideal as a baseline model (Abramovich & Ritov, 2013).

3.4.1. Assumptions of multiple linear regression

Multiple linear regression has four assumptions. First, multiple linear regression can only accurately estimate the relationship between the dependent and independent variables when the relationship is linear (Waters & Jason, 2002). To test this assumption, a scatterplot will be used to visualize the standardized residuals against the predicted dependent variable. Second, the assumption of homoscedasticity will be tested using the same scatterplot. Homoscedasticity means that the variance of errors is the same for all levels of the independent variable. Slight heteroscedasticity has little effect on significance. However, when heteroscedasticity is significantly present, it can lead to severe distortion of findings and weaken the analysis (Tabachnick & Fidell,1996). Third, the assumption of normally distributed variables is checked. This will be done using the Shapiro-Wilk Test (Shapiro & Wilk, 1965). Lastly, the low multicollinearity assumption is checked by calculating the Variance Inflation Factors (VIF-values) for all numerical independent variables. VIF-values above five can influence the model significantly, making the model less reliable (Alin, 2010).

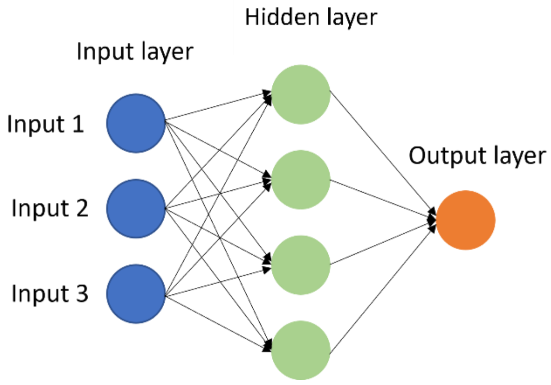
3.5. Random forest

Random forest is an algorithm proposed by Breiman (2001) which constructs multiple decision trees on different samples and picks the best-performing decision trees based on majority voting to be used in the final decision tree. Decision trees tend to overfit relatively fast and are sensitive to the data it is trained on. Random forests prevent these errors by using different samples on multiple different trees. This makes the random forest a robust algorithm (Belgiu & Drăguţ, 2016). Increasing the number of trees does not always lead to a better-performing model (Oshiro et al., 2012). The same holds for the number of randomly selected predictors. Therefore, these parameters are tuned using a random grid search.

3.6. Multilayer perceptron

Multilayer perceptron is a fully connected feedforward artificial neural network. Each multilayer perceptron consists of an input layer, at least one hidden layer, and an output layer (Ruck et al., 1990). The input layer receives the input signals, which are the independent variables and their values. The output layer converts the output of the hidden layers to a single prediction. The hidden layer converts the input using activation functions. Since it is a feedforward neural network, there is no backpropagation. Only the output of the previous layer is used as input (Ramchoun et al., 2016).

Figure 1. Example of a multilayer perceptron with one hidden layer and three inputs.



3.6.1 Activation functions and hyperparameters

Each multilayer perceptron node uses an activation function to compute the output for a given input. The most commonly used are Rectified Linear Activation (ReLU), Logistic (Sigmoid), and Hyperbolic Tangent (Tanh). Of these three functions, ReLU is the most common activation function for hidden layers in multilayer perceptron because, in contrast to the other two functions, ReLU is less susceptible to vanishing gradient (Sharma, Sharma & Athaiya, 2017). However, to confirm ReLU is the best performing activation function for this situation, all three activation functions are used in a random grid search to find the best-performing one.

The activation function for the output layer is dependent on the type of problem the model needs to solve. The final number of goals is a non-bounding numerical prediction which is best solved by an identity activation function. Since this is also the single most used output layer for regression problems, identity is used as activation function for the output layer.

Lastly, the number of hidden layers, nodes per layer, and other hyperparameters will again be chosen by the best-performing model found in the random grid search.

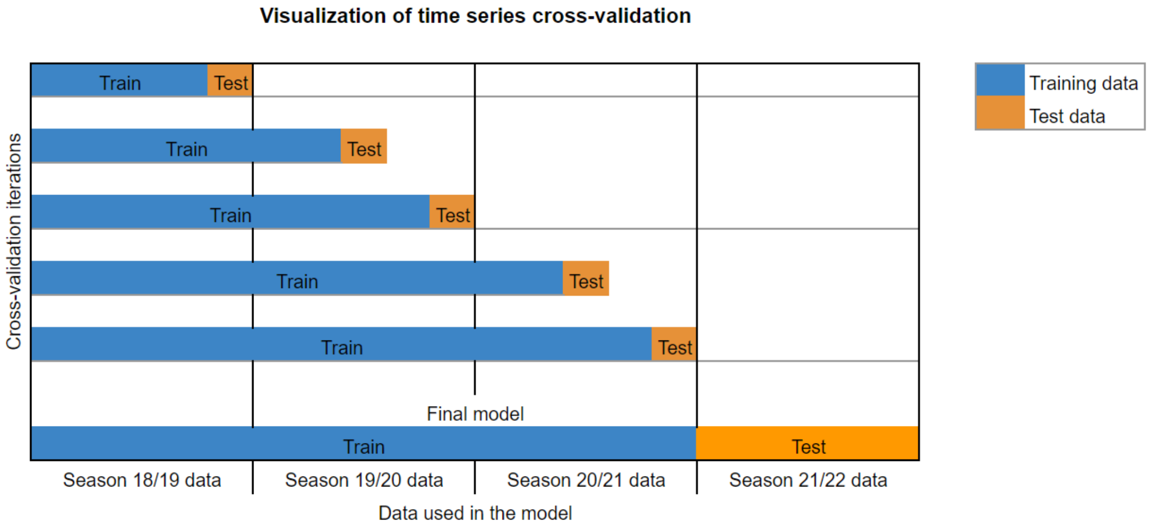
3.7. Evaluation criteria and validation

In line with the current standard in machine learning and predicting match outcomes, the dataset is split up into test and train data. The first three seasons of the 4 leagues consisting of 4083 matches (74.5%) are used to train and validate the models. The most recent season (season 21-22), consisting of 1400 matches (25.5%), is used to test the prediction accuracy of the model on new data.

3.7.1. Hyperparameter validation

Cross-validation is the standard for evaluating the performance of a model since it is more robust and prevents overfitting. This makes using cross-validation better than only splitting the data into a train and test group. However, soccer match data is time-ordered, which means that cross-validation cannot be used because this would violate one of its assumptions. Teams with a high Elo in the future have won more matches in the past, however, the model cannot know this information in real-life cases and thus, should not know this information either while training. For this reason, the time series adaptation of cross-validation is applied. A visualization of this adaptation can be found in Figure 2. Using this time series cross-validation, each model with specific hyperparameter values will run multiple times on different sub-samples of the data. The average model performance is then calculated over all iterations. In line with the research of Kumar and Roy (2018), the model with the lowest overall mean squared error (MSE) is chosen as the model with the best hyperparameters.

Figure 2. Visualization of time series cross-validation for hyperparameter tuning and final model.



3.7.2. Model evaluation

The models with the optimal hyperparameters for the multiple linear regression, random forest, and multilayer perceptron are trained on all match data of seasons 18/19 till 20/21. Subsequently, the three models predict the final number of goals for both the home and the away team for the 21/22 season. To evaluate the prediction performance of the models, the mean squared error as well as the R-squared, and R-squared adjusted will be compared to see which model produces the lowest errors and can explain the most variance in the data.

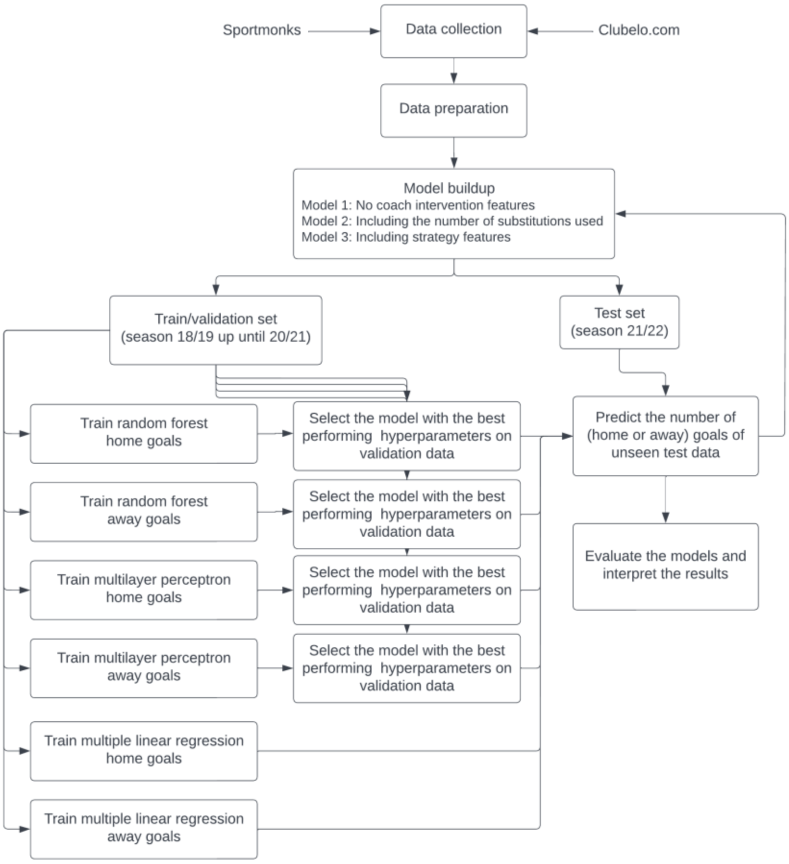
In addition for the multiple linear regression, a likelihood ratio test will be performed to calculate whether including the coach intervention features increases the model performance significantly. For the random forest and multilayer perceptron a Diebold-Mariano test will be performed to calculate whether adding the coach intervention features significantly increases the prediction performance.

Lastly, the predicted number of goals gets rounded and used to predict the match winner. If there is less than one goal difference the game will be coded as a tie. These predicted match outcomes are used to calculate the accuracy in order to compare the models with previous literature.

4. Experimental setup

Figure 3 is a visualization of the workflow of this study. In this chapter, the data collection and preparation steps are described. Furthermore, the exploratory data analysis is performed and lastly, the hardware and software used in this study are provided in detail.

Figure 3. Flowchart illustrating the steps from data collection to model evaluation


Sportmonks
Data collection
Clubelo.com
Data preparation
Model buildup
Model 1: No coach intervention features
Model 2: Including the number of substitutions used
Model 3: Including strategy features
Train/validation set
(season 18/19 up until 20/21)
Test set
(season 21/22)
Train random forest home goals
Select the model with the best performing hyperparameters on validation data
Predict the number of (home or away) goals of unseen test data
Train random forest away goals
Select the model with the best performing hyperparameters on validation data
Evaluate the models and interpret the results
Train multilayer perceptron home goals
Select the model with the best performing hyperparameters on validation data
Train multilayer perceptron away goals
Select the model with the best performing hyperparameters on validation data
Train multiple linear regression home goals
Train multiple linear regression away goals

4.1. Data collection

The data used in this study were the four most recent finished seasons from the four biggest professional European competitions. An overview of the number of matches per season and per competition can be seen in Appendix 1.1. To collect the data, two data sources were used. First, for each match, the list of events is collected via Sportmonks. This list contains the event (yellow card, red card, goal, and substitution) and the players involved. For each player involved in a substitution, the most played position of that player is retrieved. This is used to calculate the strategy change. In addition to the information retrieved from Sportmonks, the website clubelo.com is used. For each match, the Elo rating of the home and away team is scraped for the given match date.

4.2. Data preparation

In order to run the algorithms, the data must be combined in one dataset and each match must be split into 90 observations, all representing one minute of the match. In the 90 rows of the same match, all predetermined match features are the same. Next, this data is augmented with the dynamic in-game features (yellow cards, red cards, goals, and the number of substitutions used). If the first event, for example, is a yellow card for the home team in the 6th minute, the total number of yellow cards for the home team column is zero from up until row 5 and for rows 6 to 90 gets updated to 1. This updating continues until each event is added to the dataset.

For the change in strategy, the average position difference between the substituted and substitute must be calculated. The average position of both players is given as a categorical number where 1 is the keeper, 2 is a defender, 3 is a midfielder, and 4 is an attacker. The difference between both players gets added to the dataset as previous events. For example, if an attacker (4) gets substituted for a defender (2), the strategy score decreases by two from the minute the substitution is made until the 90th minute. An overview of the variables with their description can be found in Table 3.

4.3. Exploratory data analysis

A summary of the descriptive statistics of the full dataset is given in Table 5. For each variable, the mean and standard deviation is given. For the dynamic in-game features, the mean and standard deviation at 45 minutes and 90 minutes are reported.

Table 5. Descriptive statistics (N=5483)

Variables	**Mean**		**Standard deviation**	
Dependent variable				
Final number of home goals	1.445		1.312	
Final number of away goals	1.142		1.171	
Independent variable				
Predetermined match features				
Home Elo	1709.137		112.738	
Away Elo	1708.654		111.881	
Dynamic in-game features	**45 minutes**	**90 minutes**	**45 Minutes**	**90 minutes**
Home goals	0.597	1.382	0.815	1.312
Away goals	0.498	1.142	0.721	1.170
Home yellow cards	0.686	1.980	0.795	1.467
Away yellow cards	0.830	2.199	0.866	1.337
Home red cards	0.007	0.078	0.085	0.282
Away red cards	0.018	0.099	0.134	0.310
Coach interventions features				
Number of home substitutions	0.124	2.889	0.341	0.368
Number of away substitutions	0.102	2.896	0.337	0.370
Home strategy	0.001	0.136	0.228	1.442
Away strategy	0.002	0.133	0.276	1.472

From figure 4 and Table 5 it can be seen that, on average, the home team scores 1.382 goals per match. This is, on average, 0.24 goals more than the away team. This was expected due to the home advantage. In Figure 5, it can be seen that coaches make hardly any substitutions during the first half. Some coaches use their substitutions during the break, however, the number of substitutions rises quickly after the 60th minute of the match. Lastly, in Figure 6. the strategy score for both the home and away team is displayed. It can be seen that both the home and away teams' on average increase their strategy score after the halftime break, and after the 80th minute decrease their strategy score.

Figure 4. Total number of goals per minute for the home and away team

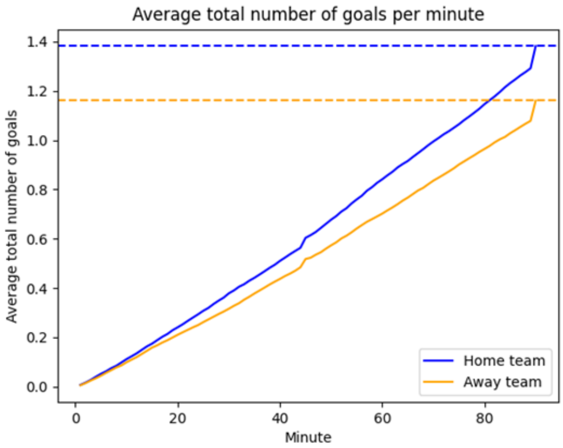


Figure 5. The cummulative number of substitutions used per minute for the home and away team.

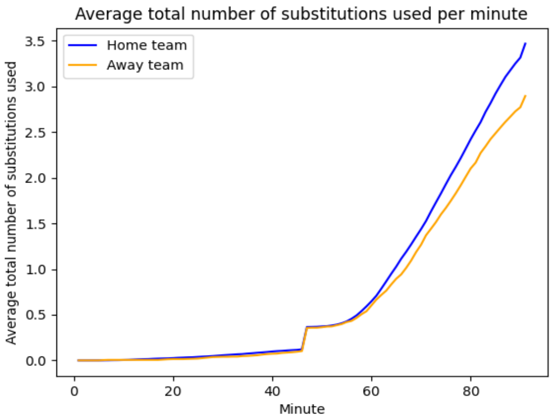
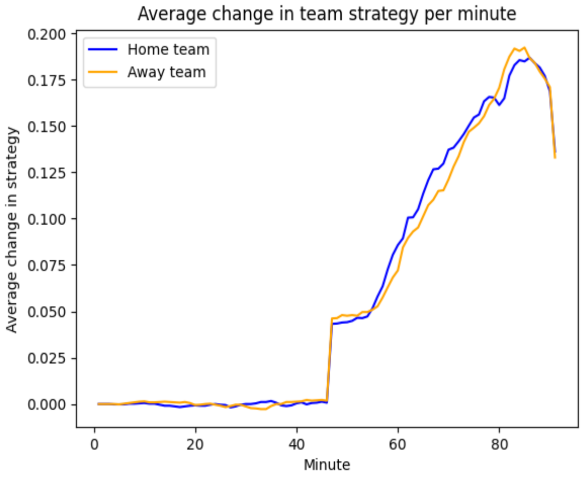


Figure 6. Average change in strategy score per minute for the home and away team.



4.4. Hardware/software used

Both data preparation and analysis are done in python 3.10. For the data collection, packages pandas, requests, and json are used. For the analysis, the additional packages, matplotlib, numpy, and sklearn were used. Everything is performed on a Microsoft Surface 4 laptop with an i7 processor and intel iris xe graphic card.

5. Results

In this section, the empirical analysis results are presented. First, the multiple linear regression baseline results are shown. Next, the overall model fit and prediction performances are shown and evaluated across the different models. Lastly, the model performances on subsamples are evaluated to see if there are notable differences in performance.

5.1. Baseline assumptions

In order to correctly interpret the results of multiple linear regression, the assumptions are tested. All assumption test results can be found in Appendix 2. First, multicollinearity of the variables is tested. The variable minute was the only variable with a VIF score above the threshold of 5 and, for that reason, is removed from the multiple linear regression model. The Shapiro-Wilk Test produced a p-value below 0.05, indicating the residuals do not follow a normal distribution. Log transforming the dependent variable did not change this finding, and neither did log transforming the independent variables. Plotting the standardized residuals against the predicted values showed there is a linear relationship between de independent variables and the dependent variables. This graph also shows there is heteroscedasticity in the data. Again, log transforming the dependent and the independent variables did not change this outcome. Since two assumptions have not been met, the findings of the multiple linear regression have to be interpreted with caution.

5.2. Model results

In this section, the predictive performance of the models are shown on the test data (season 21/22). All three models gradually increase in complexity. The first model contains no coach intervention in player lineup features. The second model includes the change in strategy feature, and the third model also contains the number of substitutions used. For each model, the optimal hyperparameters are used based on the validation data. The used hyperparameters and the chosen ones can be found in Appendix 3.

5.2.1. Comparing model prediction performance

From Table 7 it can be observed that, as expected, the mean squared error reduces the further the game progresses. Furthermore, for multiple linear regression, random forest, and multilayer perceptron, the models including all features overall had the lowest mean squared error, the highest R-squared (see Appendix 4.3), and the highest R-squared adjusted (see Appendix 4.2). The multiple linear regressions without coach intervention features have an average MSE of 0.834 goals for the home team and 0.695 for the away team. Including the cumulative number of substitutions used significantly decreases the MSE for the home team to 0.816, with a likelihood ratio of 19605 (p-value=<0.01) and decreases the away MSE to 0.690 with a likelihood ratio of 16301 (p-value=<0.01). Including the strategy change feature significantly decreased the MSE of the home goals to 0.811 with a likelihood ratio of 10002 (p-value=<0.01) and decreased the MSE of the away goals to 0.685 with a likelihood ratio of 16278 (p-value=<0.01).

The random forest without coach intervention features had a MSE of 0.792 for the home goals and 0.671 for the away goals. Adding the cumulative number of substitutions used significantly decreased the MSE for the home goals to 0.776 with a Diebold-Mariano statistic of 25.12 (p-value=<0.01), and decreased the MSE for the away goals to 0.656 with a Diebold-Mariano statistic of 15.53 (p-value=<0.001). Adding the change in strategy feature significantly decreased the MSE of the home goals to 0.767 with a Diebold-Mariano statistic of 17.48 (p-value=<0.01), and significantly decreased the MSE of the away goals to 0.647 with a Diebold-Mariano statistic of 18.54 (p-value=<0.01).

Lastly, the multilayer perceptron without coach intervention features had a MSE of 0.771 for the home goals and 0.666 for the away goals. Adding the cumulative number of substitutions used significantly decreased the MSE for the home goals to 0.762 with a Diebold-Mariano statistic of 11.16 (p-value=<0.01), and decreased the MSE for the away goals to 0.652 with a Diebold-Mariano statistic of 13.71 (p-value=<0.01). The addition of the change in strategy feature significantly decreased the MSE of the home goals to 0.751 with a Diebold-Mariano statistic of 27.29 (p-value=<0.01), and significantly decreased the MSE of the away goals to 0.638 with a Diebold-Mariano statistic of 46.65 (p-value=<0.01).

From table 7 it can be seen that the accuracy of predicting the match winner also improves by including both coach intervention features. The average accuracy of multiple linear regression without coach intervention features is 61.38% and increases to 62.56% when the coach intervention features are added. The average accuracy of the random forest without coach intervention features is 63.01%. Adding the coach intervention features increases the average accuracy to 63.57%. The average accuracy of multilayer perceptron without coach interventions is 62.79%, and this increases to 63.98% when the coach intervention features are included.

5.2.2. Comparing model techniques

Since the predictions with all features significantly outperform the other models, the models including all features are used to analyze the difference between multiple linear regression, random forest, and multilayer perceptron.

The random forest significantly outperforms the multiple linear regression at predicting the number of home goals with a 0.044 lower MSE, 0.027 higher R-squared, and a Diebold-Mariano statistic of 28.27 (p-value=<0.01). The random forest also significantly outperforms the multiple linear regression at predicting the number of away goals with a 0.039 lower MSE, 0.025 higher R-squared, and a Diebold-Mariano statistic of 32.22 (p-value=<0.01). Moreover, the random forest predicts the match winner on average 1.01% more often correctly.

The multilayer perceptron significantly outperforms both the multiple linear regression with a Diebold-Mariano statistic of 18.71 (p-value=<0.01) as well as the random forest with a Diebold-Mariano statistic of 12.58 (p-value=<0.01) at predicting the number of home goals. The multilayer perceptron also significantly outperforms multiple linear regression (Diebold-Mariano statistic= 49.29, p-value=<0.01) and the random forest (Diebold-Mariano statistic= 13.54, p-value=<0.01) at predicting the number of away goals. On average the multilayer perceptron has a 0.06 lower MSE for predicting home goals and a 0.05 lower MSE for predicting away goals compared to multiple linear regression and a 0.02 lower MSE for predicting home goals and a 0.01 lower MSE of predicting away goals compared to random forest. The average accuracy of the multilayer perceptron is 1.42% higher compared to multiple linear regression and 0.41% compared to random forest.

Table 6. Predictive performance on the test data across the models measured in mean squared error

	Multiple linear regression						Random forest						Multilayer perceptron					
	Model 1		Model 2		**All features**		Model 1		Model 2		**All features**		Model 1		Model 2		**All features**	
	Home	Away	Home	Away	**Home**	**Away**	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	**Home**	**Away**
Overall	0.8341	0.6953	0.8156	0.6902	**0.8108**	**0.6854**	0.7919	0.6712	0.7762	0.6556	**0.7667**	**0.6469**	0.7714	0.6660	0.7616	0.6516	**0.7510**	**0.6378**
90min	0.2443	0.1505	0.1435	0.1038	**0.1380**	**0.0991**	0.1467	0.1173	0.0693	0.0490	**0.0413**	**0.0272**	0.0272	0.0690	0.0392	**0.0131**	**0.0191**	0.0368
75min	0.4407	0.3248	0.3965	0.3105	**0.3917**	**0.3060**	0.3535	0.2856	0.3349	0.2666	**0.3236**	**0.2580**	0.3438	0.2910	0.3229	0.2646	**0.3189**	**0.2600**
50min	0.7210	0.6018	0.7166	0.6001	**0.7121**	**0.5956**	0.6942	0.5987	0.6903	0.5909	**0.6798**	**0.5846**	0.6988	0.5938	0.6852	0.5937	**0.6709**	**0.5810**
25min	1.1177	0.9487	1.1074	0.9399	**1.1024**	**0.9348**	1.1136	0.9369	1.1064	0.9191	**1.0999**	**0.9084**	1.0945	0.9240	1.0956	0.9178	**1.0805**	**0.8969**
0	**1.5351**	**1.3520**	1.5512	1.3688	1.5458	1.3637	1.4855	1.2779	1.4720	1.2724	**1.4665**	**1.2654**	1.4163	1.2719	1.4117	1.2381	**1.4065**	**1.2171**

Model 1 is the baseline model without coach intervention in player lineup features. Model 2 is the baseline model and the change in strategy feature included. Model 3 contains all features. In bold is the best performing model.

Table 7. Predictive performance on the test data across the models measured in accuracy

	Multiple linear regression			Random forest			Multilayer perceptron		
	Model 1	Model 2	**All features**	Model 1	Model 2	**All features**	Model 1	Model 2	**All features**
Overall	0.6138	0.6196	**0.6256**	0.63011	0.6351	**0.6357**	0.6279	0.6375	**0.6398**
90min	0.9029	0.9076	**0.9164**	0.9729	0.9914	**0.9999**	0.9700	0.9749	**0.9757**
75min	0.7443	0.7495	**0.7571**	0.7707	0.7779	**0.7793**	**0.7771**	**0.7771**	0.7764
50min	0.6236	**0.6320**	0.6307	0.6286	**0.6293**	0.6286	0.6279	0.6371	**0.6379**
25min	0.525	0.5303	**0.5464**	0.5407	0.5407	**0.5464**	0.5264	0.5436	**0.5529**
0	0.4129	0.4184	**0.4264**	0.430	**0.4407**	0.4393	0.4529	**0.465**	0.4629

Model 1 is the baseline model without coach intervention in player lineup features. Model 2 is the baseline model and the change in strategy feature included. Model 3 contains all features. In bold is the best performing model.

5.2.3. Coach interventions in player lineups

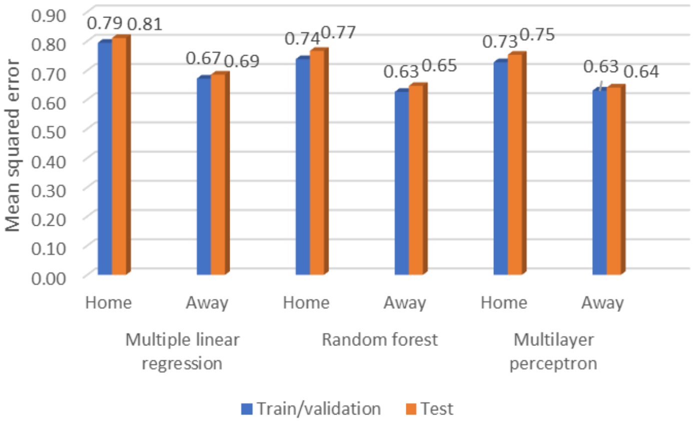
From the multiple linear regression results (Appendix 4.1) it can be observed that the number of substitutions used for both the home and away team has a negative effect on the final number of goals. This implies that when a substitution is made, on average, the average number of goals for the home team is reduced by 0.0889 (β= 0.0889, p-value =< 0.001) and by 0.0072 (β= 0.0072, p-value =< 0.001) for the away team. The same results can be observed from the random forest (home goals -0.0020, away goals -0.0015) and multilayer perceptron (home goals -0.0079, away goals -0.0011) see Appendix 5.

From the multiple linear regression results (Appendix 4.1) it can be observed that changes in the away team strategy do not have a significant effect on the final number of goals for the home (β= -0.0006, p-value =0.768) or away team (β= 0.0014, p-value =0.422). Changing the home team's strategy does have a small but significant negative effect on the final number of goals for the home (β= -0.0059, p-value =< 0.001) and away team (β= -0.0100, p-value =< 0.001). This same effect is found by the random forest (home goals -0.0055, and away goals -0.0065) see Appendix 6. The multilayer perceptron found that, if the home team plays more attacking the number of predicted home goals increases (0.0603) and the number of away goals decreases (-0.0253). If the home team change to a more defensive lineup both the predicted number of home goals (0.0050) and the predicted number of away goals (0.0253) increase.

5.3. Error analysis

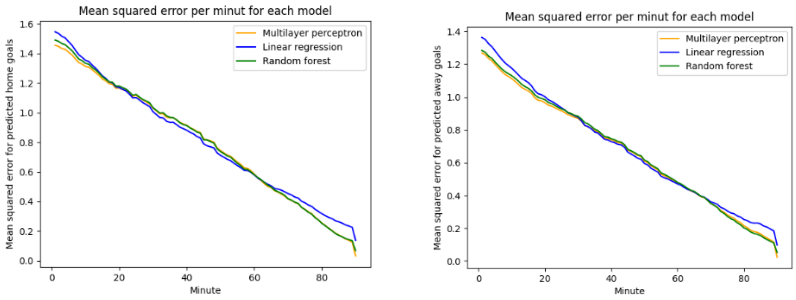
Before drawing conclusions, an error analysis is performed. Firstly, the generalization of the models is checked by comparing the prediction performance of the models on the train data compared to the test data see figure 7. Since the models are trained and validated using time series cross-validation no large under nor overfitting is found.

Figure 7. Average mean squared error for each model on the train/validation data and on the test data.



The mean squared residuals per minute of the three best-performing models are visualized in Figure 8. It can be observed that all three models predict nearly the same mean squared error from 20 to 60 minutes. From 60 minutes, the non-linear models are better at predicting the final number of goals for both the home and away team.

Figure 8. Mean squared error per minute for each model for the predicted home and away goals



In addition, the mean squared errors of the first 140 matches (10%) are compared with the mean squared errors of the last 140 matches (See appendix 7.1). From the results it can be observed that all three models performed better on the last 10% of the matches. This can be partly explained by the Elo rating, which gets more accurate at the end of a season compared to the start because most player changes happen at the start of the season. Since all three models improve relatively the same, they can still be compared using the overall measures.

In order to evaluate if the model over-predicts a particular category (home win, tie, away win) the confusion matrix of the best-performing model is visualized in Table 9. Compared to the actual winner, the model underestimates the home team to win 1.04% of the time, overestimates the game to tie 4.38% of the time, and underestimates the away team to win 4.55% of the time.

Table 9. Confusion matrix of the multilayer perceptron including all coach intervention features transformed to predict the match winner or tie.

		Predicted			
		Home	Tie	Away	**Total Actual**
Actual	Home	425	123	29	41.21%
	Tie	112	193	76	27.21%
	Away	52	115	275	31.57%
	Total predicted	42.06%	31.76%	27.19%	

Lastly, because the multiple linear regression found high multicollinearity for the feature ‘minute’ the random forest and multilayer perceptron were trained and tested without this feature. Both the random forest as well as the multilayer perceptron decreased in explained variance and had higher mean squared errors (see Appendix 7.1). This indicates that the models, including the minute feature, are preferred over the data used for the multiple linear regression.

6. Conclusion and discussion

In conclusion, the results are linked back to the research questions. Starting with the first research question.

> *RQ1. To what extent can the final number of goals (for the home and away team) be predicted using a combination of predetermined match factors and dynamic in-game features during the match?*

The most accurate model, which was the multilayer perceptron including all coach intervention features, on average had an error of 0.658 goals for predicting the number of home goals, and an average error of 0.599 goals for predicting the number of away goals. In order to compare the performance with previous literature the accuracy was calculated which was 63,98%. Compared to the study of Robberechts et al. (2021) which had an average accuracy of 74% this is lower than the current benchmark. This can be explained by the lack of meaningful dynamic in-game feature data which is further explained in the limitations section.

> *RQ1.a How accurately can an artificial neural network predict the final number of goals per match (for the home and away team) using a combination of predetermined match factors and dynamic in-game features during the match, compared to machine learning models such as linear regression and random forest?*

The artificial neural network used in this study, namely the multilayer perceptron, is significantly more accurate compared to both multiple linear regression and random forest at predicting the final number of goals a team will score. This is in line with the findings in sports like cricket (Kumar & Roy, 2018; Bhawkar & Patel, 2019), stating that multilayer perceptron is more accurate at predicting the final number of points. However, the finding is contrary to the study of Capobianco et al. (2019). They found that random forest was better at predicting the soccer match winner at halftime compared to multilayer perceptron. An explanation of this is that Capobianco et al. only used one season of data which could be too low for an artificial neural network. In this study more data is used which favors more complex models.

> *RQ2. To what extent do coach interventions in player lineups during soccer matches influence the final number of goals for the home and away team?*

Including the number of substitutions used and the strategy feature increased the explained variance and decreased the mean squared error significantly. This implies that changes in both features can significantly change the number of goals for both the home and away team. The number of substitutions used by either the home or away team had a negative relationship with the number of goals of both the home and away team. Combining this finding with previous findings that substituted players reduce the fatiguing effects of a team (Liu et al., 2019), could indicate that fatigued players allow for more errors which benefit the attacking team and thus increase the number of goals. For the strategy, a small negative relationship was found between the home strategy and the number of goals for both the home and away team.

> *RQ2.a To what extent does changing to a more defensive player lineup affect the number of predicted goals for the home and away team in a soccer match?*

If the home team changes to a more defensive lineup both the number of home and away goals significantly increase. If the away team changes to a more defensive lineup no significant effect is found.

> *RQ2.b To what extent does changing to a more attacking player lineup affect the number of predicted goals for the home and away team in a soccer match?*

If the home team changes to a more attacking player lineup the number of home and away goals significantly decrease. If the away team changes to a more attacking lineup no significant effect is found.

6.1. Empirical and societal impact

This study has shown a way of using the final number of home and away goals to build models and find effects influencing this dependent variable. By doing so it has set a benchmark of predicting the final number of home and away goals since no previous research has made these available. Furthermore, it is the first to apply an artificial neural network to predict soccer matches using live (per minute) data.

In addition, soccer clubs now have empirical data that coaches can use to make better decisions during the game. For example, a coach of a home team that has a lead now has statistical proof that the team could benefit from using a substitution to change to a more defensive player lineup to reduce the chance of a future goal being made. Moreover, coaches could use the models of this study to simulate their current game and decisions to predict the situational best statistical outcome. Coaches should keep in mind that, whether these findings also hold for other seasons or leagues has not been tested. However, given that the results from the train/validation data and unseen test data did not largely diverse indicates that findings are likely to be generalizable.

As mentioned in the introduction, consuming sports media has been linked to improved well-being, less loneliness, and higher self-esteem (Kim et al., 2017; Council of Europe, 2022). From the results it is concluded that reducing the number of substitutions used increases the number of goals per match. Therefore, reducing the allowed number of substitutions per match increases the excitement of matches, increasing the number of viewers and watch time, and ultimately enhancing the named benefits in favor of society.

6.2. Limitations and future research

One of the limitations of this research was the lack of certain in-game data resources. Previous research found an important effect of, for example, the percentage of won duals and pass accuracy for each minute of the match, and the winning chances (Robberechts et al., 2021). This data is not readily available for historical matches, and no collaboration with one of the distributors could be arranged. This lack of certain in-game data limits the accuracy of the models from this study which explain the low accuracy of 63% compared to the current benchmark of 74% set by Robberechts et al. (2021). Future research could include more relevant dynamic in-game features to more accurately predict the match outcome. In addition, this thesis had no information regarding the extra time. Every event which happened in the extra time counted as an event happing in the 45th or 90th minute.

Lastly, given the time constraints of this study, only one deep learning model was included. Future research could include other deep learning models, for example, recurrent neural networks. These less commonly used models have not shown strong performance in predicting match outcomes in other sports. However, their performance in predicting the outcomes and number of goals in soccer matches have yet to be studied.

References

Abramovich, F., & Ritov, Y. (2013). Statistical Theory: A Concise Introduction (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b14755

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374. https://doi.org/10.1002/wics.84

Anfilets, S., Bezobrazov, S., Golovko, V., & Sachenko, A. (2020). Deep multilayer neural network for predicting the winner of football matches. *International Journal of Computing*, *19*(1), 70–77. https://doi.org/10.31891/1727-6209/2020/19/1-70-77

Bailey, N., & S Clarke, S. R. (2006). Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress. Journal of Sports Science and Medicine. https://pubmed.ncbi.nlm.nih.gov/24357940/

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bhawkar, Aniket & Patel, Bhumi. (2019). Predicting the winner of Indian Premier League match: A Comparative Study. 10.13140/RG.2.2.21550.56642.

Bradley, P. S., & Noakes, T. D. (2013). Match running performance fluctuations in elite soccer: Indicative of fatigue, pacing or situational influences? *Journal of Sports Sciences*, *31*(15), 1627–1638. https://doi.org/10.1080/02640414.2013.796062

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brocas, I., & Carrillo, J. D. (2004). Do the “Three-Point Victory” and “Golden Goal” Rules Make Soccer More Exciting? Journal of Sports Economics, 5(2), 169–185. https://doi.org/10.1177/1527002503257207

Caballero, P., Garcia Rubio, J., & Ibáñez, S. J. (2017). Influence of situational variables on the U’18 soccer performance analysis (Análisis de la influencia de las variables situacionales en el rendimiento en futbol U’18). *Retos*, *32*, 224–227. https://doi.org/10.47197/retos.v0i32.56071

Capobianco, G., Di Giacomo, U., Mercaldo, F., Nardone, V., & Santone, A. (2019). Can Machine Learning Predict Soccer Match Results? Proceedings of the 11th International Conference on Agents and Artificial Intelligence. https://doi.org/10.5220/0007307504580465

Cook, R. (2001). Linear Hypothesis: Regression (Graphics). *International Encyclopedia of the Social &Amp; Behavioral Sciences*, 8888–8893. https://doi.org/10.1016/b0-08-043076-7/00455-1

Cooper, H., Deneve, K. M., & Mosteller, F. (1992). Predicting Professional Sports Game Outcomes from Intermediate Game Scores. CHANCE, 5(3–4), 18–22. https://doi.org/10.1080/09332480.1992.10554981

Council of Europe. (2022). Culture and Sport. Manual for Human Rights Education With Young People. https://www.coe.int/en/web/compass/culture-and-sport

Dawson, P., Dobson, S., & Gerrard, B. (2000). Estimating Coaching Efficiency in Professional Team Sports: Evidence from English Association Football. *Scottish Journal of Political Economy*, *47*(4), 399–421. https://doi.org/10.1111/1467-9485.00170

Dixon, M. J., & Coles, S. G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265–280. https://doi.org/10.1111/1467-9876.00065

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

Elo, A. E. (1978). The Rating of Chessplayers, Past and Present. Arco Pub.

Erickson, B. J., Chalmers, P. N., Bush-Joseph, C. A., & Romeo, A. A. (2016). Predicting and Preventing Injury in Major League Baseball. *The American Journal of Orthopedics*, *45*(3), 152–156. https://pubmed.ncbi.nlm.nih.gov/26991568/

Fahey-Gilmour, J., Dawson, B., Peeling, P., Heasman, J., & Rogalski, B. (2019). Multifactorial analysis of factors influencing elite australian football match outcomes: a machine learning approach. International Journal of Computer Science in Sport, 18(3), 100–124. https://doi.org/10.2478/ijcss-2019-0020

Ganguly, S., Frank, N.: The problem with win probability. In: Proc. of the 12th MIT Sloan Sports Analytics Conf. (2018)

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14–15), 2627–2636. https://doi.org/10.1016/s1352-2310(97)00447-0

Gill, P. S. (2000). Late-Game Reversals in Professional Basketball, Football, and Hockey. The American Statistician, 54(2), 94–99. https://doi.org/10.1080/00031305.2000.10474518

Hills, S. P., Barwood, M. J., Radcliffe, J. N., Cooke, C. B., Kilduff, L. P., Cook, C. J., & Russell, M. (2018). Profiling the Responses of Soccer Substitutes: A Review of Current Literature. *Sports Medicine*, *48*(10), 2255–2269. https://doi.org/10.1007/s40279-018-0962-9

Huang, M. L., & Lin, Y. J. (2020). Regression Tree Model for Predicting Game Scores for the Golden State Warriors in the National Basketball Association. Symmetry, 12(5), 835. https://doi.org/10.3390/sym12050835

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. International Journal of Forecasting, 26(3), 460–470. https://doi.org/10.1016/j.ijforecast.2009.10.002

Igiri, C. P., & Nwachukwu, E. O. (2014). An Improved Prediction System for Football a Match Result. IOSR Journal of Engineering, 04(12), 12–020. https://doi.org/10.9790/3021-04124012020

Jia, M., Zhao, Y., Chang, F., Zhang, B., & Yoshigoe, K. (2020). A Random Forest Regression Model Predicting the Winners of Summer Olympic Events. *Proceedings of the 2020 2nd International Conference on Big Data Engineering*. https://doi.org/10.1145/3404512.3404513

Joseph, A., Fenton, N., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), 544–553. https://doi.org/10.1016/j.knosys.2006.04.011

Kayhan, V. O., & Watkins, A. (2019). Predicting the point spread in professional basketball in real time: a data snapshot approach. Journal of Business Analytics, 2(1), 63–73. https://doi.org/10.1080/2573234x.2019.1625730

Kumar, S., & Roy, S. (2018). Score Prediction and Player Classification Model in the Game of Cricket Using Machine Learning. International Journal of Scientific & Engineering Research, 9(8).

Liu, H., Wang, L., Huang, G., Zhang, H., & Mao, W. (2019). Activity profiles of full-match and substitution players in the 2018 FIFA World Cup. *European Journal of Sport Science*, *20*(5), 599–605. https://doi.org/10.1080/17461391.2019.1659420

Lorenzo-Martínez, M., Rein, R., Garnica-Caparrós, M., Memmert, D., & Rey, E. (2020). The Effect of Substitutions on Team Tactical Behavior in Professional Soccer. *Research Quarterly for Exercise and Sport*, *93*(2), 301–309. https://doi.org/10.1080/02701367.2020.1828563

Maher, M. J. (1982). Modelling association football scores. Statistica Neerlandica, 36(3), 109–118. https://doi.org/10.1111/j.1467-9574.1982.tb00782.x

McCabe, A., & Trevathan, J. (2008). Artificial Intelligence in Sports Prediction. Fifth International Conference on Information Technology: New Generations (Itng 2008). https://doi.org/10.1109/itng.2008.203

Merritt, S., & Clauset, A. (2014). Scoring dynamics across professional team sports: tempo, balance and predictability. EPJ Data Science, 3(1). https://doi.org/10.1140/epjds29

Misener, L., Taks, M., Chalip, L., & Green, B. C. (2015). The elusive “trickle-down effect” of sport events: assumptions and missed opportunities. *Managing Sport and Leisure*, *20*(2), 135–156. https://doi.org/10.1080/23750472.2015.1010278

Osborne, J. W. (n.d.). *Four assumptions of multiple regression that researchers should always test*. ScholarWorks@UMass Amherst. Retrieved November 2, 2022, from https://scholarworks.umass.edu/pare/vol8/iss1/2/

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*, 154–168. https://doi.org/10.1007/978-3-642-31537-4_13

Paola, M. D., & Scoppa, V. (2011). The Effects of Managerial Turnover. *Journal of Sports Economics*, *13*(2), 152–168. https://doi.org/10.1177/1527002511402155

Pollard, R. (2008). Home Advantage in Football: A Current Review of an Unsolved Puzzle. *The Open Sports Sciences Journal*, *1*(1), 12–14. https://doi.org/10.2174/1875399x00801010012

Prasetio, D., & Harlili, D. (2016). Predicting football match results with logistic regression. 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA). https://doi.org/10.1109/icaicta.2016.7803111

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer Perceptron: Architecture Optimization and Training. *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(1), 26. https://doi.org/10.9781/ijimai.2016.415

Ramsey, A., World Cup medal honour for Sir Alf . https://citas.in/frases/1833970-alf-ramsey-never-change-a-winning-team/. 26 June 2009

Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. (2017). Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). IOP Conference Series: Materials Science and Engineering, 226, 012099. https://doi.org/10.1088/1757-899x/226/1/012099

Rein, R., Raabe, D., & Memmert, D. (2017). “Which pass is better?” Novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science*, *55*, 172–181. https://doi.org/10.1016/j.humov.2017.07.010

Robbe Ruck, D., Rogers, S., & Kabrisky, M. (1990). Feature Selection Using a Multilayer Perceptron. *Journal of Neural Network Computing*, *2*(2), 40–48.

rechts, P., Van Haaren, J., & Davis, J. (2021). A Bayesian Approach to In-Game Win Probability in Soccer. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3512-3521).

Rudrapal, D., Srivastava, S., & Singh, S. (2019). A Deep Learning Approach to Predict Football Match Result. In Advances in Intelligent Systems and Computing. (990th ed., pp. 93–99). https://doi.org/10.1007/978-981-13-8676-3_9

Schalkoff, R. (1992). Pattern Recognition: Statistical, Structural and Neural Approaches, John Wiley & Sons. *Inc, New York*.

Shaprio, S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *towards data science*, *6*(12), 310-316.

Shirley, K. (2007). Markov model for basketball. In Proceedings of the New EnglandSymposium for Statistics in Sports, Boston, MA.

Silva, R. M., & Swartz, T. B. (2016). Analysis of substitution times in soccer. Journal of Quantitative Analysis in Sports, 12(3). https://doi.org/10.1515/jqas-2015-0114

Stefani, Raymond. (1977). Football and basketball predictions using least squares. IEEE Transactions on Systems, Man, and Cybernetics. 7. 117-121.

Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. International Journal of Forecasting, 28(2), 532–542. https://doi.org/10.1016/j.ijforecast.2011.01.004

Stübinger, J., Mangold, B., & Knoll, J. (2019). Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. Applied Sciences, 10(1), 46. https://doi.org/10.3390/app10010046

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics . Northridge. *Cal.: Harper Collins*.

Taud, H., & Mas, J. (2017). Multilayer Perceptron (MLP). *Geomatic Approaches for Modeling Land Change Scenarios*, 451–455. https://doi.org/10.1007/978-3-319-60801-3_27

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA Game Result Prediction Using Feature Analysis and Machine Learning. Annals of Data Science, 6(1), 103–116. https://doi.org/10.1007/s40745-018-00189-x

The Business Research Company. (2022, January). Sports Market Size 2022 And Growth Analysis. https://www.thebusinessresearchcompany.com/report/sports-global-market-report

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Waters, E., & Jason, O. W. (2002). Four Assumptions of Multiple Regression That Researchers Should Always Test. *Practical Assessment, Research and Evaluation*, *8*(2), 2. https://doi.org/10.7275/r222-hv23

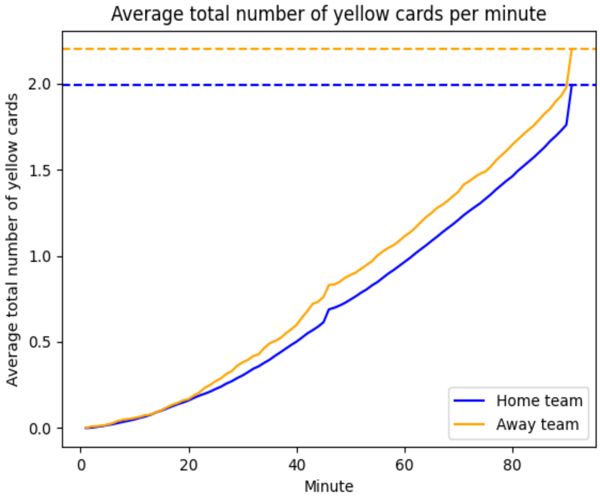
Appendix

1. Descriptive statistics

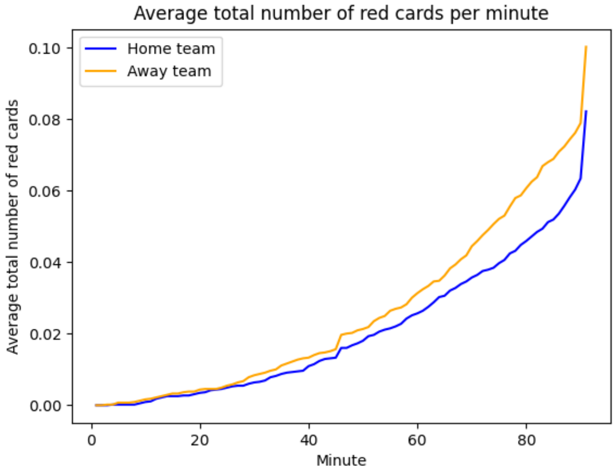
1.1 Number of matches per competition and season

Competition	Season 18/19	Season 19/20	Season 20/21	Season 21/22
UK Premier League	380	288	380	371
IT Serie A	380	233	380	379
ES La Liga	380	380	380	366
GE Bundesliga	289	307	306	284
Total per season	1,429	1,208	1,446	1,400

1.2 The cumulative number of yellow cards per minute for the home and away team.



1.3 The cumulative number of red cards per minute for the home and away team.



2. Assumptions test results of multiple linear regression

2.1 VIF scores of features used in multiple linear regression

Feature	VIF score
Competition Bundesliga	1.5427
Competition LaLiga	1.6116
Competition Premier League	1.6932
season 19/20 (dummy season)	1.5884
Season 20/21 (dummy season)	1.6624
Season 21/22 (dummy season)	1.6203
Minutes played	8.9935
Difference in Elo (Home - Away)	1.1193
Cumulative home goals	2.0628
Cumulative away goals	1.8774
Cumulative home red cards	1.0704
Cumulative away red cards	1.0890
Cumulative home yellow cards	2.3657
Cumulative away yellow cards	2.5990
Cumulative number of substitutions used home	3.6866
Cumulative number of substitutions used away	3.7979
Change in strategy home	1.0829
Change in strategy away	1.0870

2.2 Breusch-Pagan test for heteroscedasticity of the multiple linear regression

	Model 1		Model 2		Model 3	
	Home	Away	Home	Away	Home	Away
Lagrange multiplier statistic	32605.2560	26621.3175	32605.2560	26621.3175	32605.2560	26621.3175
P-value	0.001	0.001	0.001	0.001	0.001	0.001
F-value	2053.5722	1655.1955	2053.5722	1655.1955	2053.5722	1655.1955
F test p-value	0.001	0.001	0.001	0.001	0.001	0.001

2.3 Shapiro-Wilk Test for normality of residuals of the multiple linear regression

2.3.1 Dependent variable final number of goals

	Model 1		Model 2		Model 3	
Shapiro-Wilk normality	Home	Away	Home	Away	Home	Away
Test statistic	0.9968	0.9339	0.9970	0.9339	0.9339	0.9970
P-value	0.001	0.001	0.010	0.001	0.001	0.001

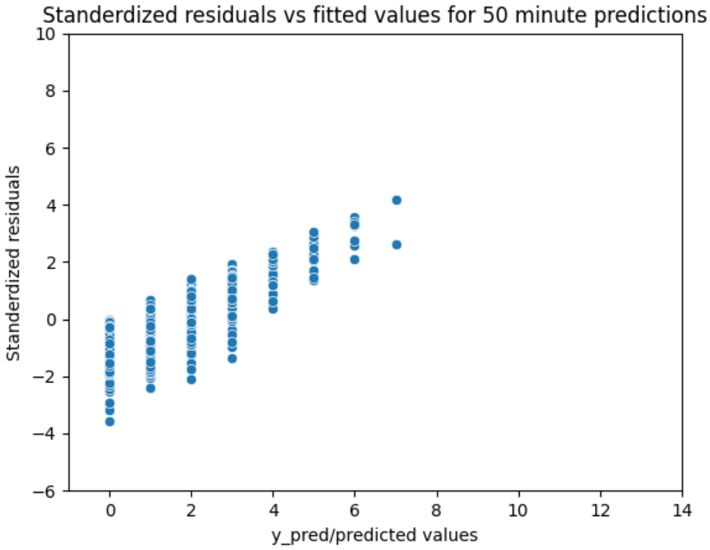
2.3.2 Dependent variable log transformed

	Model 1		Model 2		Model 3	
Shapiro-Wilk normality	Home	Away	Home	Away	Home	Away
Test statistic	0.9930	0.9955	0.9914	0.9982	0.9915	0.9981
P-value	0.001	0.001	0.001	0.013	0.001	0.015

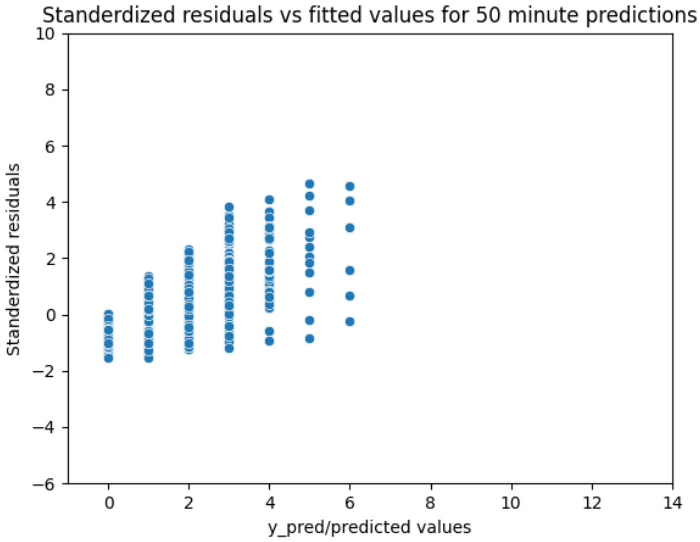
2.3.3 Log transformed dependent variable as well as independent variables

	Model 1		Model 2		Model 3	
Shapiro-Wilk normality	Home	Away	Home	Away	Home	Away
Test statistic	0.9956	0.9955	0.9940	0.9983	0.9940	0.9983
P-value	0.001	0.001	0.001	0.019	0.001	0.017

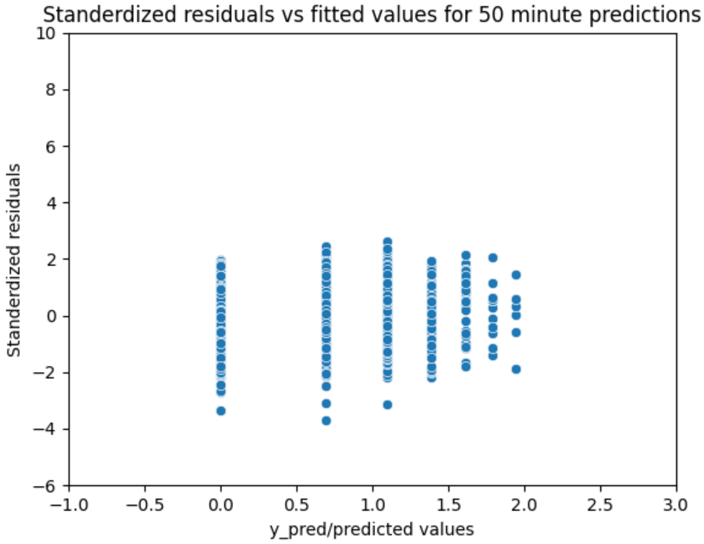
2.4 Standardized residuals of the full home multiple linear regression model plotted against the predicted values for the 50th minute.



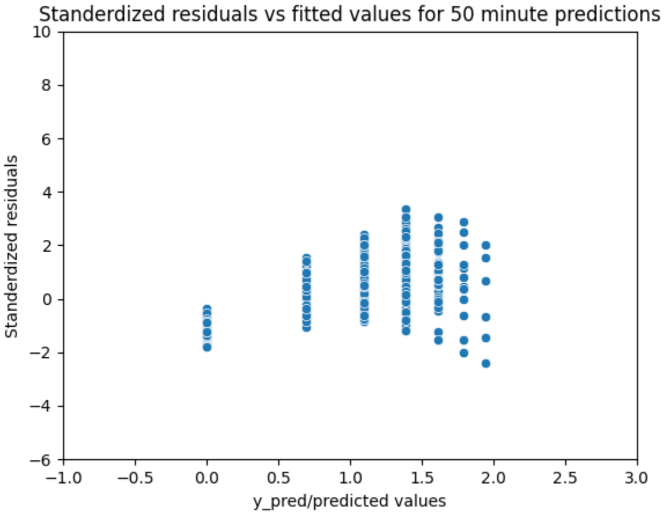
2.5 Standardized residuals of the full away multiple linear regression model plotted against the predicted values for the 50th minute.



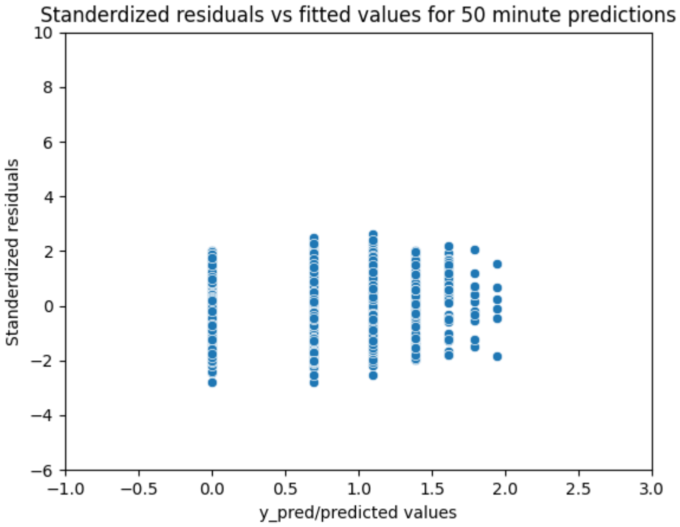
2.6 Standardized residuals of the full home multiple linear regression model plotted against the predicted values for the 50th minute with log-transformed dependent variable.



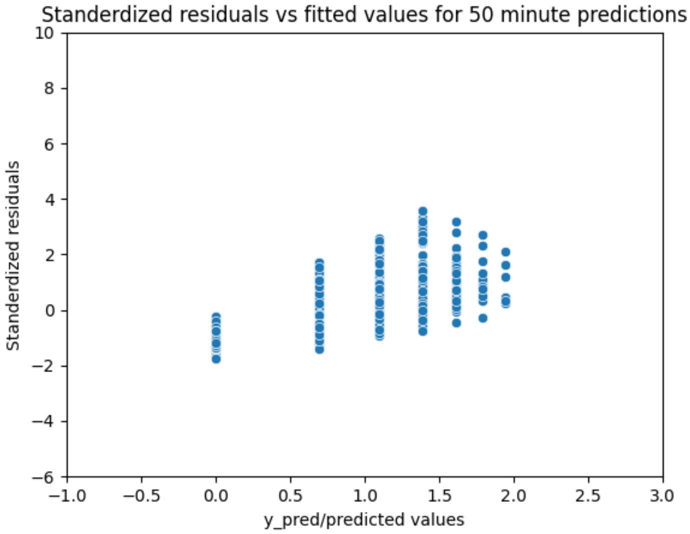
2.7 Standardized residuals of the full away multiple linear regression model plotted against the predicted values for the 50th minute with log-transformed dependent variable.



2.8 Standardized residuals of the full home multiple linear regression model plotted against the predicted values for the 50th minute with log-transformed dependent variable as well as independent variables.



2.9 Standardized residuals of the full away multiple linear regression model plotted against the predicted values for the 50th minute with log-transformed dependent variable as well as independent variables.



3. Hyperparameter tuning

3.1 Random grid search used and best-performing hyperparameters for random forest

Random forest	Model 1		Model 2		Model 3	
	Home	Away	Home	Away	Home	Away
Max depth [2,5,8,10,12,15,20]	6	6	6	6	6	6
Max features [2,4,6,8,10,12,15]	6	8	8	8	12	12
Min samples leaf [2,5,8,10,12,15,20]	5	5	5	5	5	8
Min samples split [1,2,3,4,8,10,12]	8	10	8	8	10	10
Number estimators [100,200,400,1000]	200	200	200	200	200	200

3.1 Random grid search used and best-performing hyperparameters for multilayer perceptron

Multilayer perceptron	Model 1		Model 2		Model 3	
	Home	Away	Home	Away	Home	Away
Number of hidden layers [1,2,3,4,5]	1	1	2	2	2	2
Number of perceptrons per layer [2,4,5,6,8,9,10,12,14,16,20,24,28,32,48,64,128]	10	6	8,4	9,4	9,2	9,4
Activation [ReLu, Identity, Logistic, Tanh]	ReLu	ReLu	ReLu	ReLu	ReLu	ReLu
maximum iterations [50,70,100,200,400,1000]	400	200	400	400	400	400

4. Model results

4.1 Multiple linear regression results

	Model 1		Model 2		Model 3	
Predictors	**Home**	**Away**	**Home**	**Away**	**Home**	**Away**
Constant	1.0322**	0.8494**	0.9898**	0.8138**	0.9900**	0.8141**
Difference in Elo (Home - Away)	0.0015**	-0.0011**	0.0015**	-0.0011**	0.0015**	-0.0011**
Cumulative home goals	0.8671**	-0.0763**	0.9191**	-0.0286**	0.9186**	-0.0295**
Cumulative away goals	-0.0488**	0.9063**	0.0008	0.9426**	0.0014	0.9439**
Cumulative home red cards	-0.2710**	0.1755**	-0.2132**	0.2217**	-0.2164**	0.2162**
Cumulative away red cards	0.1747**	-0.2104**	0.2354**	-0.1593**	0.2365**	-0.1561**
Cumulative home yellow cards	-0.1166**	-0.1079**	-0.0613**	-0.0634**	-0.0613**	-0.0635**
Cumulative away yellow cards	-0.1471**	-0.1050**	-0.0905**	-0.0574**	-0.0904**	-0.0573**
Cumulative number of substitutions used home			-0.1030**	-0.0651**	-0.1025**	-0.0642**
Cumulative number of substitutions used away			-0.0752**	-0.0817**	-0.0755**	-0.0823**
Change in strategy home					-0.0059**	-0.0095**
Change in strategy away					-0.0001	0.0020
Competition Bundesliga	0.1056**		0.1950**	0.1259**	0.1950**	0.1259**
Competition LaLiga	0.0808**		0.1096**	0.0920**	0.1094**	0.0918**
Competition Premier League	-0.0483**		0.0524**	-0.0276**	0.0524**	-0.0276**
Season 19/20 (dummy season)	0.0394**		-0.0118**	0.0396**	-0.0119**	0.0395**
Season 20/21 (dummy season)	0.0093**		-0.0377**	0.0455**	-0.0380**	0.0451**
Season 21/22 (dummy season)	0.0217**		0.0219**	0.0587**	0.0218**	0.0586*
Model summaries						
Degrees of freedom	13	13	15	15	17	17
F-statistic	3.439e+04	3.453e+04	3.176e+04	3.150e+04	28030	27770
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
AIC	1.304e+06	1.218e+06	1.289e+06	1.206e+06	1.275+06	1.200+06
R2	0.4790	0.4980	0.4889	0.5026	0.4937	0.5075
R2 adjusted	0.4788	0.4976	0.4836	0.4999	0.4933	0.5047

4.2 R-squared adjusted of all models

	Multiple linear regression						Random forest						Multi layer perceptron					
	Model 1		Model 2		**All features**		Model 1		Model 2		**All features**		Model 1		Model 2		**All features**	
	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away
Overal	0.4788	0.4976	0.4836	0.4999	**0.4933**	**0.5047**	0.5051	0.5150	0.5149	0.5263	**0.5209**	**0.5326**	0.5179	0.5187	0.524	0.5292	**0.5307**	**0.5391**
90min	0.8936	0.9216	0.8952	0.9101	**0.9137**	**0.9284**	0.9083	0.9153	0.9567	0.9646	**0.9742**	**0.9803**	0.983	0.9501	0.9755	**0.9905**	**0.9881**	0.9734
75min	0.7496	0.7783	0.7403	0.7637	**0.7552**	**0.7789**	0.7791	0.7937	0.7907	0.8074	**0.7978**	**0.8135**	0.7851	0.7898	0.7982	0.8088	**0.8007**	**0.8121**
50min	0.5515	0.5654	0.5442	0.5587	**0.555**	**0.5696**	0.5661	0.5673	0.5686	0.573	**0.5752**	**0.5775**	0.5633	0.5709	0.5718	0.571	**0.5807**	**0.5802**
25min	0.3071	0.3255	0.3049	0.3179	**0.311**	**0.3245**	0.3041	0.323	0.3085	0.3359	**0.3126**	**0.3436**	0.316	0.3323	0.3153	0.3368	**0.3247**	**0.3519**
0	0.0948	**0.0841**	0.033	0.0143	0.0339	0.0146	0.0716	0.0766	0.0801	0.0806	**0.0835**	**0.0856**	0.1149	0.081	0.1178	0.1054	**0.1210**	**0.1206**

Model 1 is the baseline model without coach intervention in player lineup features. Model 2 is the baseline model and the change in strategy feature included. Model 3 contains all features. In bold is the best performing model.

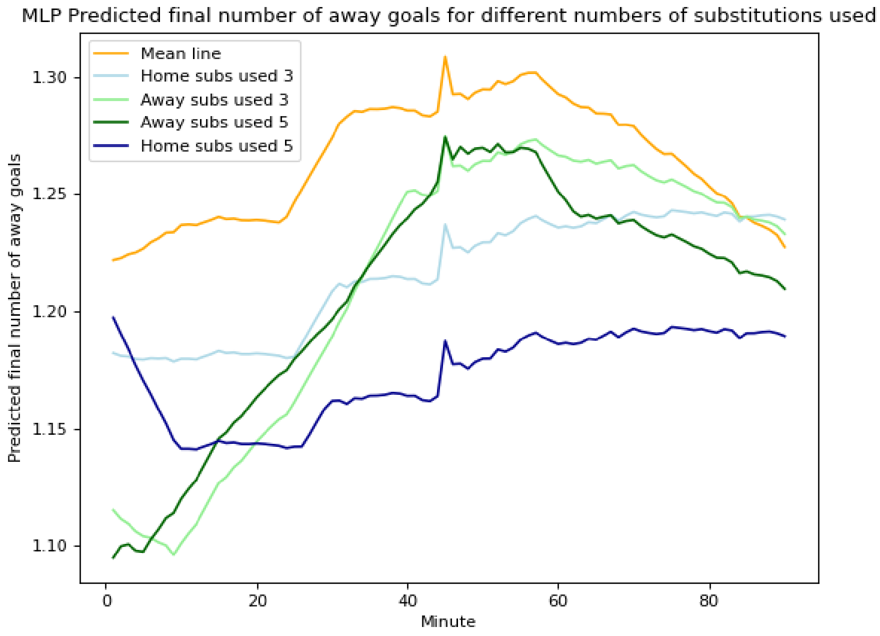
4.3 R-squared of all models

R2	Multiple linear regression						Random forest						Multi layer perceptron					
	Model 1		Model 2		**All features**		Model 1		Model 2		**All features**		Model 1		Model 2		**All features**	
	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away	Home	Away
Overal	0.479	0.498	0.4889	0.5026	**0.4937**	**0.5075**	0.5051	0.5150	0.5149	0.5264	**0.5209**	**0.5326**	0.5201	0.5215	0.5257	0.5292	**0.5314**	**0.5392**
90min	0.8473	0.8912	0.9138	0.9312	**0.9229**	**0.9401**	0.9551	0.9578	0.9812	0.9837	**0.9935**	**0.9950**	0.983	0.9515	0.9920	**0.9941**	**0.9928**	0.9889
75min	0.7246	0.7653	0.7477	0.7724	**0.7552**	**0.7799**	0.7810	0.7957	0.7907	0.8074	**0.7978**	**0.8135**	0.7906	0.7972	0.8002	0.8089	**0.8010**	**0.8127**
50min	0.5494	0.5652	0.5575	0.5671	**0.5628**	**0.5725**	0.5673	0.5674	0.5686	0.5735	**0.5752**	**0.5778**	0.5709	0.5739	0.5732	0.5712	**0.5816**	**0.5805**
25min	0.3015	0.3145	0.3106	0.3282	**0.3136**	**0.3315**	0.3055	0.3243	0.3085	0.3361	**0.3128**	**0.3436**	0.3174	0.3335	0.3159	0.3368	**0.3256**	**0.3520**
0	**0.0407**	**0.0231**	0.0945	0.0823	**0.0956**	0.0829	0.1020	0.101	0.1038	0.1023	**0.1060**	**0.1045**	0.1159	0.0813	0.1196	0.1071	**0.1222**	**0.1206**

Model 1 is the baseline model without coach intervention in player lineup features. Model 2 is the baseline model and the change in strategy feature included. Model 3 contains all features. In bold is the best performing model.

5. Non-linear full model predictions for different numbers of substitutions used.

5.1 Visualisation of full multilayer perceptron predicting the number of away goals for different numbers of substitutions used by the home and away team.



5.2 All variables having the mean value of the variable at the given minute.

	Number of substitutions used		Random forest		Multilayer perceptron	
Minute	Home	Away	Home goals	Away goals	Home goals	Away goals
45	0.121	0.115	1.840	1.660	1.410	1.337
60	0.708	0.770	1.488	1.411	1.443	1.310
70	1.526	1.588	1.353	1.311	1.489	1.240
80	2.517	2.559	1.221	1.187	1.502	1.162
90	3.469	3.469	1.150	1.140	1.454	1.079
Average number of predicted goals :			1.3030	1.2622	1.472	1.1977

5.3 Both home and away reduced the substitution by 50%

	Number of substitutions used		Random forest		Multilayer perceptron	
Minute	Home	Away	Home goals	Away goals	Home goals	Away goals
45	0.060	0.057	1.840	1.66	1.406	1.331
60	0.35	0.385	1.501	1.448	1.457	1.344
70	0.763	0.794	1.368	1.326	1.535	1.278
80	1.258	1.279	1.260	1.241	1.587	1.276
90	1.734	1.734	1.156	1.166	1.652	1.238
Average number of predicted goals :			1.3212	1.2952	1.5577	1.284
Change in the predicted number of goals, compared to the average :			18,25	33	85,75	86,25

5.4 Both home and away increase the substitutions used by 50%

	Number of substitutions		Random forest		Multilayer perceptron	
used						
Minute	Home	Away	Home goals	Away goals	Home goals	Away goals
45	0.182	0.173	1.830	1.640	1.414	1.343
60	1.062	1.155	1.468	1.401	1.449	1.307
70	2.29	2.382	1.343	1.301	1.403	1.232
80	3.777	3.84	1.200	1.167	1.364	1.148
90	5.204	5.205	1.120	1.120	1.357	1.060
Average number of predicted goals :			1282,75	1247,25	1393,25	1186,75
Change in the predicted number of goals, compared to the average :			-20,25	-15	-78,75	-11

5.5 Only the home team has decreased the substitutions used by 50%

	Number of substitutions used		Random forest		Multilayer perceptron	
Minute	Home	Away	Home goals	Away goals	Home goals	Away goals
45	0.060	0.115	1.840	1.660	1.41	1.332
60	0.35	0.770	1.513	1.422	1.469	1.317
70	0.763	1.588	1.351	1.318	1.481	1.254
80	1.258	2.559	1.220	1.206	1.557	1.186
90	1.734	3.469	1.155	1.149	1.585	1.212
Average number of predicted goals :			1.3097	1.2737	1.5230	1.2422
Change in the predicted number of goals, compared to the average :			6,75	11,5	51	44,5

5.6 Only the away team has decreased the substitutions used by 50%

	Number of substitutions used		Random forest		Multilayer perceptron	
Minute	Home	Away	Home goals	Away goals	Home goals	Away goals
45	0.121	0.057	1.840	1.660	1.417	1.337
60	0.708	0.385	1.478	1.442	1.432	1.327
70	1.526	0.794	1.373	1.326	1.494	1.233
80	2.517	1.279	1.262	1.230	1.512	1.181
90	3.469	1.734	1.151	1.160	1.522	1.195
Average number of predicted goals :			1.3160	1.2895	1.4900	1.2340
Change in the predicted number of goals, compared to the average :			13	27,25	18	36,25

6. Non-linear full model predictions for different strategy scores.

6.1 All variables having the mean value of the variable at the given minute.

	Strategy score		Random forest		Multilayer perceptron	
Minute	Home strategy	Away strategy	Home goals	Away goals	Home goals	Away goals
45	0.001	0.002	1.84	1.66	1.410	1.337
60	0.089	0.084	1.488	1.411	1.443	1.310
70	0.138	0.128	1.353	1.311	1.489	1.240
80	0.165	0.181	1.221	1.187	1.502	1.162
90	0.136	0.133	1.150	1.140	1.455	1.079
Average number of predicted goals :			1.3030	1.2622	1.4722	1.1977

6.2 Both home and away play more attacking (strategy +2)

	Strategy score		Random forest		Multilayer perceptron	
Minute	Home strategy	Away strategy	Home goals	Away goals	Home goals	Away goals
45	2.001	2.002	1.820	1.662	1.546	1.403
60	2.089	2.084	1.499	1.403	1.579	1.288
70	2.138	2.128	1.369	1.302	1.560	1.218
80	2.165	2.181	1.222	1.173	1.505	1.140
90	2.136	2.133	1.148	1.127	1.457	1.057
Average number of predicted goals :			1.3095	1.2512	1.5252	1.1757
Change in the predicted number of goals, compared to the average :			6,5	-11	53	-22

6.3 Both home and away play more defensive (strategy -2)

	Strategy score		Random forest		Multilayer perceptron	
Minute	Home strategy	Away strategy	Home goals	Away goals	Home goals	Away goals
45	-1.999	-1.998	1.840	1.660	1.569	1.149
60	-1.911	-1.916	1.488	1.408	1.370	1.170
70	-1.862	-1.872	1.353	1.310	1.353	1.235
80	-1.835	-1.819	1.221	1.187	1.409	1.184
90	-1.864	-1.867	1.150	1.140	1.452	1.101
Average number of predicted goals :			1.3030	1.2612	1.3960	1.1725
Change in the predicted number of goals, compared to the average :			0	-1	-76,25	-25,25

6.4 Only the home team plays more attacking (strategy +2)

	Strategy score		Random forest		Multilayer perceptron	
Minute	Home	Away	Home goals	Away goals	Home goals	Away goals
45	2.001	0.002	1.832	1.667	1.482	1.318
60	2.089	0.084	1.482	1.408	1.475	1.285
70	2.138	0.128	1.348	1.304	1.521	1.215
80	2.165	0.181	1.215	1.179	1.577	1.137
90	2.136	0.133	1.145	1.132	1.557	1.054
Average number of predicted goals :			1297,5	1255,75	1532,5	1172,75
Change in the predicted number of goals, compared to the average :			-5,5	-6,5	60,25	-25

6.5 Only the away team plays more attacking (strategy +2)

	Strategy score		Random forest		Multilayer perceptron	
Minute	Home strategy	Away strategy	Home goals	Away goals	Home goals	Away goals
45	0.001	2.002	1.820	1.661	1.511	1.407
60	0.089	2.084	1.495	1.412	1.505	1.304
70	0.138	2.128	1.369	1.315	1.530	1.233
80	0.165	2.181	1.222	1.188	1.485	1.155
90	0.136	2.133	1.148	1.141	1.438	1.072
Average number of predicted goals :			1308,5	1264	1489,5	1191
Change in the predicted number of goals, compared to the average :			5,5	1,75	17,25	-6,75

6.6 Only the away team plays more defensive (strategy -2)

	Strategy score		Random forest		Multilayer perceptron	
Minute	Home strategy	Away strategy	Home goals	Away goals	Home goals	Away goals
45	0.001	-1.998	1.840	1.66	1.369	1.267
60	0.089	-1.916	1.488	1.408	1.402	1.288
70	0.138	-1.872	1.353	1.310	1.448	1.247
80	0.165	-1.819	1.221	1.187	1.504	1.169
90	0.136	-1.867	1.150	1.140	1.472	1.086
Average number of predicted goals :			1303	1261,25	1456,5	1197,5
Change in the predicted number of goals, compared to the average :			0	-1	-15,75	-0,25

7. Error analysis

7.1 Mean squared errors of the predictions for the first 10% of the 21/22 season matches compared to the last 10% of the matches.

	Multiple linear regression		Random forest		Multilayer perceptron	
	Home	Away	Home	Away	Home	Away
First 10%	0,8053	0,7164	0,7266	0,6886	0,7264	0,6771
Last 10%	0,7564	0,6337	0,7033	0,5877	0,6818	0,5767
Difference	-0,0489	-0,0827	-0,0233	-0,1009	-0,0446	-0,1004

7.2 Model results of the multilayer perceptron and random forest without minute feature

	R-squared		R-squared adjusted		Mean squared error	
	Home	Away	Home	Away	Home	Away
Multilayer perceptron	0.5057	0.5131	0.5055	0.5093	0.7912	0.6791
Random forest	0.4978	0.5143	0.4976	0.5140	0.8039	0.6725