ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

# Excel bookmakers' betting odds

## Predicting soccer match outcomes using player characteristics, team ratings and bookmaker odds

Author: P.E. (Pepijn) van der Leije (455987) Date final version: 23 June 2022

Supervisor: N.W. (Nick) Koning Second Assessor:

*MSc Thesis Econometrics & Management Science*

*Business Analytics & Quantitative Marketing*

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

## Abstract

We predict the outcome of soccer matches using three machine learning models: random forest, stochastic gradient boosting, and Learning with Subset Stacking model from Birbil, Yildirim, Gökalp and Akyüz (2021). To evaluate the performance of the different models we apply four metrics: the root mean squared error; the median absolute deviation; the accuracy; and the average return on investment with betting. Using data of 14.062 soccer matches from eight European soccer leagues from the period July 2016 to March 2022, we analyse 36 input features related to player characteristics, team ratings and bookmaker odds. We determine that the stochastic gradient boosting predicts the soccer match outcomes best according to the four used metrics, followed by the Learning with Subset Stacking model, and that both models are capable to obtain an average positive return on investment of respectively 1,49% and 0,58% per match with betting. Moreover, we find statistical evidence that both models obtain a 95% confidence interval for the average return on investment of respectively [-4,17%; 7,13%] and [-5,03%; 6,25%] through bootstrapping.

# Contents

# 1 Introduction

All kinds of sport researchers compete to construct the most accurate model for predicting the outcome of sport events. For example, a coach wants to be able to predict as accurately as possible what the outcome of a match will be given the potential line-up and tactics, so he can choose the line-up and tactics that maximize his chance of winning. The need of an accurate model also applies to individual athletes. They want to know what kind of training they should do to ensure the chance of performance is maximized. Furthermore, gamblers and bookmakers require the most accurate model to achieve the highest positive return on investment. However, predicting the outcome of sport events is a demanding task. It depends on the athletes/teams performance during a match, and is being influenced by many uncertain factors such as players qualities, home ground advantage, referee decisions, weather conditions and factors which are inherently unpredictable, perhaps best called 'luck'. Due to these factors, sports (and especially soccer) turn out to be a perfect environment to study the applicability of existing prediction methods or develop new methods to be transferred to other fields of predicting (Wunderlich and Memment, 2018).

In this study, we predict the outcome of soccer matches through machine learning models. We introduce the Learning with Subset Stacking (LESS) from Birbil et al. (2021) in the field of soccer predictions and include bootstrapping statistics to find out how reliable and valid our performance metrics are (Freedman, 1981). To realise this, we pose the following research question:

- *To what extent are various machine learning models capable of predicting the outcome of soccer matches?*

Applying data of 14.062 soccer matches from eight European soccer leagues from the period July 2016 to March 2022, we analyse 36 input features related to FIFA player characteristics, *FiveThirthyEight* team ratings and *BET365* betting odds. For evaluation purposes, we randomly split the data in two subsets. The training set consists of 90% of the data, and the test set contains the remaining 10%. To examine the outcome of soccer matches, we implement a 10-fold cross-validation and a grid search to optimize the parameters. Subsequently, we apply three naïve approaches and three machine learning models: the *HOME-approach* indicates the home team always wins; the *BET-approach* indicates the team with the lowest betting odds always wins; the *FiveThirtyEight-approach* indicates the team with the highest team rating always wins; random forest which is the best performing model to predict the outcome of soccer matches according to

Stübinger, Mangold and Knoll (2020); stochastic gradient boosting which is the best performing model to predict the outcome of soccer matches according to Geurkink, Boone, Verstockt and Bourgois (2021); and LESS model from Birbil et al. (2021) which is a machine learning algorithm designed for populations where the relation between the input variables and the output variable exhibits a heterogeneous behavior across the predictor space. The algorithm starts with generating random subsets followed by training a local predictor for each subset. Then those predictors are combined to yield an overall predictor.

To evaluate the performance of the different models we apply four metrics: the root mean squared error, the median absolute deviation, the accuracy, and the average return on investment. The first two metrics measure the model performance based on the different number of goals scored by the home team and away team. In addition, the accuracy is the number of correctly predicted match outcomes out of all match outcomes, where we classify (home team wins; draw; away team wins) the outcomes based on the different number of goals scored by the home team and away team. Furthermore, the average return on investment is a ratio that compares the gain or loss from an investment relative to its cost. We consider two applications to achieve a positive return on investment with betting. Usually a soccer match bet indicates whether one expects a team to win, draw or lose its forthcoming game, which depends on the difference in the number of goals scored by the two opposing teams. Eventually, the goal of gamblers is to correctly predict a match event. Based on the outcome predictions of our applied models, we implement two trading strategies to obtain a positive return on investment: places the same amount of money on each match; and increase the amount of investment if the match prediction outcome probability increases. So, we pose the following sub-research question:

- *To what extent are various trading strategies capable of achieving a positive return on investment by applying our prediction outcomes and betting odds from the bookmaker?*

We find that the stochastic gradient boosting predicts the soccer match outcomes best according to the four performance metrics, followed by the Learning with Subset Stacking model, and that both models are capable to obtain an average positive return on investment of respectively 1,49% and 0,58% per match by using our first trading strategy where we invest the same amount of money each match. Moreover, we find statistical evidence that both models obtain a 95% confidence interval for the average return on investment of respectively [-4,17%; 7,13%] and [-5,03%; 6,25%] through bootstrapping.

# 2 Literature Review

## 2.1 Predicting the outcome of soccer matches

The first research in the field of soccer match predictions was executed by Reep and Benjamin (1968). They provided statistical evidence that the number of goal chances increases the probability to win a match. Since then, many studies have been conducted into soccer match prediction outcomes. According to the reviews of Horvat and Job (2020), Wunderlich and Memmert (2021) and Hubáček, Ondřej, Šourek and Železnŷ (2022) soccer match outcome predictions are divided into two main categories, namely human judgement and quantitative models. Moveover, each main category can be further divided into two subcategories.

Human judgement includes all forecasts that are exclusively or predominantly driven by human decisions. The first subcategory covers individual human judgement made by single persons, for example in an experimental environment or published in the media. The second subcategory of human judgement covers sources of collaborative human judgement where the forecast arises from an interaction between various persons, for example in on community based websites or social media.

Quantitative models include all forecasts exclusively or predominantly based on mathematical models or statistical methods and are divided into two subcategories. Forecasting based on external ratings/rankings refers to forecasting from ratings/rankings that are not part of the model itself and thus were not explicitly designed for this purpose. The other subcategory includes internal ratings/rankings that are part of the model itself and thus were explicitly designed for forecasting purposes like a single strength or multiple strength parameter of a team or player.

In this study, we focus only on quantitative models and omit human judgement. A further explanation is given below.

### 2.1.1 Through external ratings/rankings models

An extensively applied approach to predict the outcome of soccer matches are external rating models, introduced by Elo (1978) and originally applied to predict the outcome of chess games. After rating and ordering the teams' qualities based on a few (sometimes one) historical variables, external rating models are used to distinguish between team and assume that the highest-ranking team wins. For example, Hvattum and Arntzen (2010) examined the value of assigning ratings to soccer teams based on their past performance in order to predict the match outcome, where they applied the Elo ratings in ordered logit regression models. The Elo based methods were compared to six benchmark prediction methods, and they were found to be significantly better than all of the other methods, in terms of observed loss. Moreover, Wunderlich and Memmert (2018) elaborated the approach by combining an Elo rating model and the information from betting odds. The novel betting odds based Elo model is shown to outperform classic Elo models, demonstrating that betting odds prior to a match contain more relevant information than the result of the match itself. A further development in the field of external ratings was executed by Arntzen and Hvattum (2021), where they predict the soccer match outcomes by including team ratings and player ratings. They concluded that forecasts made when using both team ratings and player ratings as covariates are significantly better than those based on only one of the ratings.

### 2.1.2 Through internal ratings/rankings

The domain of score-based soccer modelling has generally been dominated by approaches where the past goals scored per team are assumed to follow a particular parametric probability distribution. Maher (1982), one of the founders in this research, came up with a bivariate poisson model and a double poisson model. Through his first model he introduced the notion of teams' defensive and attacking qualities and how to use them for predicting the outcome of soccer matches.

In extension to this foundation, there are several statisticians who further improved the score-based soccer model. For example, Dixon and Coles (1997) amplified Maher's research by introducing the dependency of home team advantage, which effectively increases the probabilities of low-scoring draws. Furthermore, Rue and Salvesen (2000) created, in contrast to Maher, a time variant model where they applied a brownian motion to bundle the teams' qualities in sequential rounds. A development on the studies mentioned above was done by Crowder, Dixon, Ledford and Robinson (2002). They improved the accuracy and the computational complexity by applying an autoregressive model for updating the teams' qualities. Moreover, Owen (2011) evolved the double poisson model of Maher by adding a random walk to model of the teams' qualities, and concluded improvements on similar work, in relation to the estimation of a parameter in the model.

In recent years, Koopman and Lit (2015) created a time dynamics into the bivariate poisson model using a state-space representation, and found out that the dependency between scores had little effect on the forecast performance. Afterwards, this conclusion was supported by Ley, Wiele and Eetvelde (2019). They applied ten different strength-based statistical models to model soccer match outcomes, where eventually the bivariate poisson model performed best. Subsequently, the bivariate poisson model performed also the best in the research of Koopman and Lit (2019), where they created a bivariate poisson, an ordered probit and a skellam model wherein the teams qualities were transformed by a time series model.

Beside the poisson approaches, machine learning models have become remarkably favored in the last decades. One of the first studies in this field has been executed by McCabe and Travathan (2008). They applied an artificial neural network trained with back propagation and conjugative-gradient descent to predict the outcomes in the Australian Soccer League and English Premier League and obtained the best accuracy of 75,4%. Moreover, Hucaljuk and Rakipović (2011) applied various machine learning models to predict Champions' League group stage games outcomes, where they used three different datasets. As result the logitboost (68,8%) and neural network (68,8%) achieved the highest accuracy and outperformed the naïve bayes (56,3%), bayesian network (56,3%), K-nearest neighbor (62,5%) and random forest (65,6%).

Buursma (2011) proposed several machine learning models to predict outcomes in the Dutch Soccer League based on the last fifteen years. All input features were based on the last twenty team games, and the linear regression obtained the best accuracy of 55,05%, compared to the logistics regression (54,98%), decision trees (54,57%), logitboost (54,62%), bayesnet (54,55%) and naïve bayes (54,43%). In addition, Owramipur, Eskandarian and Mozneb (2013) applied a bayesian network model implemented in a NETICA software and predicting outcomes of team FC Barcelona in Spain Soccer League with the accuracy of 92%. Furthermore, Igiri and Nwachukwu (2014) predicted English Premier League outcomes, and after optimization and feature weighting the accuracy of the artificial neural network and linear regression increased to 85% and 93%, respectively.

Shin and Gasparyan (2014) created a model to predict soccer match outcomes that makes use of only virtual data collected from the video game FIFA 2015, where 33 players characteristics were used. They concluded that the video game industry, in the attempt to simulate a realistic experience, has inadvertently collected very accurate data which can be used to solve problems in the real world, and achieved an accuracy of 83%. In addition, Tax and Joustra (2015) created a self-made dataset from public sources, consisting of thirteen seasons of Dutch Eredivisie, which included for example whether teams played in a lower league the season before, whether a new coach was employed and whether a player was injured. After comparing multiple machine learning models, the highest prediction accuracy of 54,7% was achieved by using a combination of principal component analysis and a multilayer perceptron classifier.

Prasetio and Harlili (2016) build a logistic regression model to predict match results of Barclays Premier League season, and received an accuracy of 69,51%. The teams offenses and defences were used as input features, and it was found that the significant variables were home defence and away defence. Furthermore, Teli, Zaveri and Shinde (2018) achieved the highest accuracy from the logistic regression (71,63%) compared to the random forest (69,90%), artificial neural network (69,20%), support vector machine (66,95%) and naïve bayes (63,57%), by predicting the outcome of soccer matches using twelve features related to game performance, and records for all teams against each other during the last five seasons.

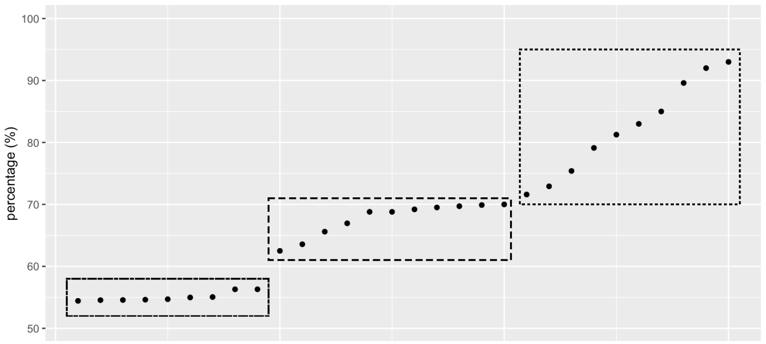
Danisik, Lacko and Farkas (2018) applied a long-short-term-memory neural networks model for outcome prediction in different soccer leagues. Incorporating player-level, obtained from the FIFA video games, and match history data, a total of 139 features were included. Eventually, it was stated that betting odds and tactics could be included in the future. Another study which agrees this statement is from Odachowski and Grekow (2012). They conducted a machine learning model to predict the soccer match outcome solely based on betting odds features and achieved a prediction accuracy of 70%.

One of the most recent studies was done by Stübinger, Mangold and Knoll (2020). They applied player characteristics (e.g., age, weight, height) and player skills (e.g., ball control, dribbling, crossing) from the FIFA video game, and predicted the match outcome for the five biggest soccer leagues (England, France, Germany, Italy, Spain) and the corresponding second leagues. In total 40 features for each player for 47.856 soccer matches were used as input variables. In their prediction model the dependent variable is the difference in goals between the home team and the away team. To describe the relation of dependent and independent variables, the underlying functions are created by the following machine learning approaches: random forest, stochastic gradient boosting and support vector machine. Altogether 21.698 home victories (45,3%), 12.895 draws (27%) and 13.253 away victories (27,7%) were achieved. These results clearly demonstrate the home advantage. Therefore, one of their benchmark strategy was that the home team always wins. In addition, their second benchmark strategy was to bet on the event with the lowest odd value from online bookmaker *BET365*, which is one of the leading betting providers with around 23 million customers. Since this is the most presumable outcome. Eventually, the best accuracy of 81,26% was achieved using random forest algorithm, followed by gradient boosting algorithm (79,12%), linear regression (72,92%) and support vector machine (69,71%). They concluded that tree-based machine learning algorithms seem to capture the information in the data best.

Geurkink, Boone, Verstockt and Bourgois (2021) aimed to identify the strongest predictive variables of the outcome of soccer matches in the highest Belgian soccer division by applying machine learning models, where they applied variance inflation factor (threshold of five) and borutashap to avoid multicollinearity and reduce dimensionality. With a total of thirteen features, they achieved an accuracy of 89,6% using stochastic gradient boosting and showed as well as Stübinger et al. (2020) that boosting is one of the best performing models for prediction the outcome of soccer matches.

Figure 1 shows the scatter plot of the soccer match outcome prediction accuracies of the studies mentioned above. It is notable that the applied match characteristics and the out-of-sample/in-sample predictions differ across studies, which results into big differences in accuracy. Therefore, we split the observations in three parts. The most upper right predictions (ca [70%-95%] accuracy) include data that is be available after the match, such as shots on target from the penalty box, and they applied the same data for training and testing purposes. It is not surprising this results into higher accuracies compared to out-of-sample predictions (ca [62%-70%] accuracy). Moreover, applying only match characteristics that are available before the match started and out-of-sample predictions result into the lower left class (ca [54%-56%]), and our study participates in this last sector.

Figure 1: Scatter plot of soccer match prediction accuracies



*Notes.* A scatter plot of the soccer match outcome prediction accuracies of previous studies is shown. Since the applied match characteristics and the out-of-sample/in-sample predictions differ across studies, we split the accuracies observations in three parts: in-sample predictions (ca [70%-95%] accuracy) including data that is be available after the match, out-of-sample predictions (ca [62%-70%] accuracy) including data that is be available after the match, and out-of-sample predictions (ca [54%-56%] accuracy) including data that is be available before the match.

### 2.2 Tradings strategy

The goal of Stübinger et al. (2020) was to generate positive returns over time by betting based on a strategy which successfully identifies overvalued betting odds. After predicting the difference in goals between the home team and away team by means of a random forest, they defined their trading strategy as the follows: if difference in goals between the home team and away team is larger than 2, they bet on "home team wins", if difference in goals between the home team and away team is smaller than -2, they bet on "away team wins", and if difference in goals between the home team and away team is in between 2 and -2 they do not forecast a clear victory of home team or away team, and no trading is executed. Eventually, their machine learning model and trading strategy achieved a positive return of 1.58% per match. However, it should be mentioned that their obtained results are in-sample.

## 3 Data

To perform this study, we combine several public sources containing characteristics on player qualities on the one hand and ratings/forecasts on the other hand. In the next section, we detail the content of these data sets as well as describe preprocessing steps that we undertook before predicting the outcome of soccer matches.

### 3.1 Data characteristics

Our dataset, provided by Football-Data[1] (to acquire *BET365* betting odds), Kaggle[2] (to acquire FIFA player characteristics) and FiveThirtyEight[3] (to acquire *FiveThirtyEight* team ratings), consists of 14.062 soccer matches from eight European soccer leagues from the period July 2016 to March 2022, as shown in Table 1. The dataset contains all kind of match characteristics such as the team playing against each other, the match outcome and the date the match is played. It should be mentioned that the difference in the number of matches per division arises, besides the diversity of seasons, from the number of teams in each league and the number of played matches in the current season '21/'22.

Table 1: Overview of soccer division data

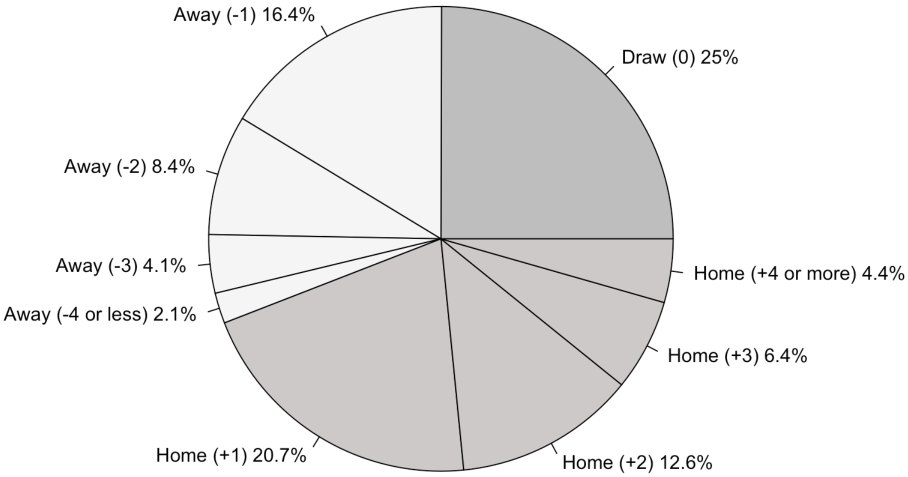
| Division | Seasons | Matches |
|---|---|---|
| Belgium Jupiler Pro League | '18/'19 - '21/'22 | 1055 |
| Dutch Eredivisie | '17/'18 - '21/'22 | 1381 |
| English Premier League | '16/'17 - '21/'22 | 2183 |
| French Ligue 1 | '16/'17 - '21/'22 | 1869 |
| French Ligue 2 | '17/'18 - '21/'22 | 1704 |
| German Bundesliga | '16/'17 - '21/'22 | 1763 |
| Italy Serie A | '16/'17 - '21/'22 | 1928 |
| Spanish La Liga | '16/'17 - '21/'22 | 2179 |

*Notes.* The divisions, seasons and number of matches are shown from the dataset. Here, eight European soccer leagues from July 2016 to March 2022 result in 14.062 soccer matches.

This study aims to predict the outcome of soccer matches through three naïve approaches and three machine learning models. To do so, we first create our dependent variable $y$: the number of goals scored by the home team minus the number of goals scored by the away team. To clarify this matter a positive value of $y$ indicates the home team wins, a negative value indicates the away team wins, and is the difference in goals is 0 the match had ended into a draw. Figure 2 shows the percentage distribution of $y$, where we examine that in 44,1% of the matches the home team wins. In addition, we notice that a deviation close to draw (e.g. Away (-1)/Home (+1)) is more likely than a large goal difference (e.g. Away (-4 or less)/Home (+4 or more).

[1] https://www.football-data.co.uk/data.php
[2] https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset/version/2
[3] https://projects.fivethirtyeight.com/soccer-predictions/

Figure 2: Percentage distribution of $y$



*Notes.* The percentage distribution of $y$, the number of goals scored by the home team minus the number of goals scored by the away team, is shown.

Furthermore, the dataset contains 36 input features related to FIFA player characteristics, *FiveThirtyEight* team ratings and *BET365* betting odds for each soccer match. It is important to mention that all these 36 characteristics are known before the match starts. An explanation of each source of information including a table overview is given in the following subsections.

### 3.1.1 FIFA player characteristics

FIFA player characteristics is information that reflects the quality and the feature of a player's technical skills, behaviours and performance on the pitch. These scores are rated first by Electronic Arts' data reviewers who are made up of coaches, professional scouts, and a lot of season ticket holders from around the world. Afterwards, Electronic Arts editors go through these data to review them and apply their own feedback. Shin et al. (2014) provided statistical evidence for the reliability of FIFA data.

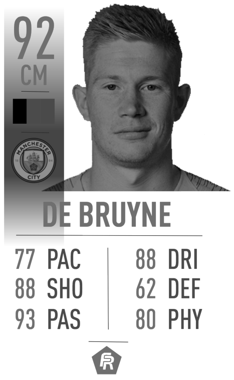


Figure 3: FIFA player characteristics of Kevin De Bruyne

Figure 3 shows the FIFA's player characteristics of soccer player Kevin de Bruyne. Here the name, position, country flag and soccer club are given. In addition, FIFA includes seven major attributes all within a range from 1-100: pace (PAC), shooting (SHO), passing (PAS), dribbling (DRI), defending (DEF) physicality (PHY), and an OVERALL score of 92 is generated. Besides these seven dominant ratings, we also include the stamina, mentality and vision of each player (all within a range from 1-100). Furthermore, the skills, which have a range from 1-5, and the weekly wage of each player are included. Thus, in total we obtain from each soccer player for each season twelve FIFA player characteristics. Nevertheless, we do not have information about which players are lined-up each match. Therefore, we take the average score of the twelve FIFA ratings from the original (the line-up preferred by the coach) ten field players for each club each season. Subsequently, the goalkeeper has other major attributes than field players, such as diving and reflexes, and we also include the overall score of the original (preferred by the coach) goalkeeper. In conclusion, we obtain for each club for each season thirteen attributes which indicate their strength and weaknesses.

### 3.1.2 *FiveThirtyEight* rating

Our second data source provide information about ESPN's Soccer Power Index ratings (SPI), originally created by *FiveThirtyEight* editor-in-chief Nate Silver in 2009 for rating international soccer teams. We include these SPI ratings, match outcome probabilities and expected number of goals as well.

**SPI ratings**

The SPI ratings are estimates of a team's overall strength within a range from 1-100. Before a season starts, the team's SPI ratings are based on two factors: its ratings at the end of the previous season (67% influence), and its market value as calculated by Transfermarkt[4] (33% influence), a website that assigns a monetary value to each player, based on what they would fetch in a transfer.

[4] https://www.transfermarkt.com

As a season plays out, the team's ratings are adjusted after every match based on its performance in that match and the strength of its opponent. Unlike with the Elo rating system, SPI ratings do not necessarily improve whenever it wins a match, if a soccer team still performs worse than expected, its ratings can decline.

**Goals expectation and match outcome probabilities**

Given two teams' SPI ratings and the assumption that goal-scoring in soccer follows a poisson process, the likelihood of the number of goals each team will score is obtained. Based on these likelihoods, including league-specific home-field advantage and the importance of the match to each team, the expected number of goals of each team are realized. After comparing the expected number of goals of each team the probability of the final score is captured which generates home/draw/away match outcome probabilities.

### 3.1.3 *BET365* betting odds

We apply corresponding betting odds from the online bookmaker *BET365*, one of the leading betting provides in the world. In bets, odds represent the ratio between the amounts wagered by parties to a bet or wager. To do so, a bookmaker creates their opening odds based on their expertise. Odachowski et al. (2012) and Danisik et al. (2018) both provide statistical evidence that implementing odd values is valuable for prediction the outcome of soccer matches, and therefore we include this information in our study.

Table 2 shows the counterplay between the probability outcomes and the resulting betting odds for Watford vs Tottemham 01-01-2017, where the estimated probability can be obtained by 1 divided by the betting odd (and conversely). The betting odds never reflect the true probability of an event occurring. There is always a profit margin added by the bookmaker since the sum of estimated probabilities is larger than one. According to our data the odds vary between 1.02 and 41, and not surprisingly the bookmaker is familiar with the well-known home advantage since the average odds of a home win (2.804) are significantly below the average odds for an away win (4.544).

An overview of the dependent variable $y$ and the input features of FIFA player characteristics, the *FiveThirtyEight* team ratings and *BET365* betting odds can be found in Table 3.

Table 2: Counterplay between the probability and betting odds

| | Estimated probability | Betting odd |
|---|---|---|
| Watford wins (Home): | 18,18% | 5,50 |
| Draw: | 26,67% | 3,37 |
| Tottenham wins (Away): | 61,72% | 1,62 |

*Notes.* The counterplay between the probability outcomes and the betting odds are shown for Watford vs Tottemham 01-01-2017, where the estimated probability can be obtained by 1 divided by the betting odd (and conversely).

Table 3: Descriptive statistics of the dependent variable and input features

| | Min. | Quart. 1 | Median | Quart. 3 | Max. | Mean |
|---|---|---|---|---|---|---|
| *Dependent variable* | | | | | | |
| y | -13 | -1 | 0 | 1 | 9 | 0.3036 |
| Scored goals home | 0.000 | 1.000 | 1.000 | 2.000 | 9.000 | 1.541 |
| Scored goals away | 0.000 | 0.000 | 1.000 | 2.000 | 13.000 | 1.238 |
| *FIFA player characteristics* | | | | | | |
| Overall | 62.90 | 70.70 | 75.10 | 78.30 | 88.10 | 74.71 |
| Pace | 55.90 | 66.70 | 69.70 | 72.60 | 81.10 | 69.64 |
| Shooting | 43.30 | 55.00 | 59.60 | 63.00 | 74.50 | 59.18 |
| Passing | 49.10 | 61.70 | 65.90 | 69.50 | 80.20 | 65.82 |
| Dribbling | 55.10 | 65.90 | 69.90 | 73.40 | 84.90 | 69.71 |
| Defending | 43.60 | 55.70 | 59.80 | 63.70 | 72.90 | 59.73 |
| Physic | 58.60 | 68.30 | 70.60 | 72.90 | 81.10 | 70.47 |
| Stamina | 57.60 | 70.50 | 73.20 | 76.20 | 86.50 | 73.12 |
| Mentality | 48.60 | 64.30 | 68.00 | 71.50 | 80.70 | 67.70 |
| Vision | 41.70 | 59.00 | 63.50 | 68.00 | 81.10 | 63.67 |
| Skills | 2.000 | 2.600 | 2.800 | 3.000 | 3.800 | 2.796 |
| Weekly wage | 1340 | 10100 | 24350 | 43300 | 319000 | 38603 |
| Goalkeeper | 57.00 | 72.00 | 77.00 | 82.00 | 92.00 | 76.51 |
| *FiveThirtyEight Rating* | | | | | | |
| SPI home | 12.60 | 47.97 | 61.60 | 72.09 | 96.57 | 59.44 |
| SPI away | 12.51 | 47.93 | 61.49 | 72.00 | 96.69 | 59.38 |
| 538 home | 0.0281 | 0.3334 | 0.4371 | 0.5509 | 0.9775 | 0.4486 |
| 538 away | 0.0037 | 0.1946 | 0.2838 | 0.3893 | 0.8992 | 0.3045 |
| 538 draw | 0.0188 | 0.2271 | 0.2577 | 0.2794 | 0.4129 | 0.2469 |
| 538 goals home | 0.410 | 1.230 | 1.460 | 1.760 | 4.900 | 1.549 |
| 538 goals away | 0.200 | 0.880 | 1.120 | 1.430 | 4.010 | 1.199 |
| *BET365 betting odds* | | | | | | |
| B365H | 1.020 | 1.700 | 2.200 | 3.100 | 26.000 | 2.804 |
| B365D | 2.000 | 3.300 | 3.600 | 4.200 | 21.000 | 4.046 |
| B365A | 1.050 | 2.370 | 3.390 | 5.000 | 41.000 | 4.544 |

*Notes.* From top to bottom: The dependent variable $y$ derives from the number of goals scored by the home team minus the number of goals scored by the away team. The 36 input features, which are known before a the match start, are related to FIFA player characteristics, *FiveThirtyEight* team rating and *BET365* betting odds.

# 4 Methodology

This section is split into three parts. First, we describe the naïve approaches and machine learning models that we use to explain and predict outcome of soccer matches. Afterwards, we consider two applications on the prediction outcomes to achieve a positive return on investment with betting. Subsequently, we describe the applied bootstrapping statistics.

## 4.1 Predicting the outcome of soccer matches

The goal of this study is to predict the outcome of soccer matches using our dataset of 14.062 matches and 36 input variables. For evaluation purposes, we randomly split the data into two subsets. The training set consists of 90% of the data, 12.657 matches, and the test set contains the remaining 10%, 1.405 matches. According to the recommendation of Birbil et al. (2021) we scale the input data to a zero mean and one standard deviation, since scaling the data makes it easier for a model to learn due to less variation between observations. To examine the outcome of soccer matches, we implement a 10-fold cross-validation and a grid search to optimize the parameters. The metric used to determine the best settings is, by default, the root mean squared error. Subsequently, we apply the following approaches/models: *HOME-approach*, *BET-approach*, *538-approach*, random forest (RF), stochastic gradient boosting (GBM) and LESS model (LESS). An explanation of the six approaches/models, including their advantages and disadvantages, is provided below. The regression codes of the random forest and the stochastic gradient boosting are programmed in R with the package 'caret' (Kuhn, 2009). The LESS model is coded in Python with the package 'LessRegressor' (Birbil, 2021).

Our dependent variable $y$ ($\in \mathbb{Z}$) indicates the difference between the number of goals scored by the home team and away team. A positive integer denotes home team wins, a negative integer denotes away team wins and 0 points at a draw. However, the predicted outcomes of $y$ are real numbers ($\in \mathbb{R}$) since we apply regression techniques. Therefore, we create boundaries to understand how we need to classify the outcomes. To do so, we apply an in-sample forecast with our trainset and assume that the highest 44.1% outcomes (5.583 matches) indicate the home team wins, the lowest 30,9% outcomes (3.915 matches) denote the away team wins, and all value in between point out draws, just as in the original dataset. Table 4 shows the classification boundaries of 'home', 'draw' and 'away' for each machine learning model. Eventually, we use these boundaries to classify the predicted outcomes of the testset.

Table 4: Classification boundaries

| | AWAY | DRAW | HOME |
|---|---|---|---|
| RF | ⟨-∞, -0.236] | ⟨ -0.236, 0.433 ⟩ | [0.433, ∞⟩ |
| GBM | ⟨-∞, -0.055] | ⟨ -0.055, 0.395 ⟩ | [0.395, ∞⟩ |
| LESS | ⟨-∞, -0.023] | ⟨ -0.023, 0.383 ⟩ | [0.383, ∞⟩ |

*Notes.* The classification boundaries of the random forest, stochastic gradient boosting and LESS model for away, draw and home are shown. We obtained them after applying an in-sample forecast with our trainset and assuming that the highest 5.583 values (44,1%) indicate the home team wins, the lowest 3.915 values (30,9%) denote the away team wins, and all value in between point out draws.

To evaluate the performance of the different approaches/models we apply four metrics: the root mean squared error (RMSE), the median absolute deviation (MAD), the accuracy and the average return on investment. The first two metrics obtain the model performance based on the difference between the number of goals scored by the home team and away team. The root mean squared error is the square root of the mean of the square of all of the error,

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\bar{y_i} - y_i\right)^2},$$

whereas the median absolute deviation is a more robust measure of the variability,

$$MAD = median|\bar{y_i} - y_i|.$$

The accuracy is the number of correctly predicted match outcomes dividend by all match outcomes, where we classify (home team wins ‘Home’; draw ‘Draw’; away team wins ‘Away’) the outcomes based on the different number of goals scored by the home team and away team. Here, the correct predictions are Home-Home, Draw-Draw and Away-Away. The other outcome options, which are incorrect, are Home-Draw, Home-Away, Dram-Home, Draw-Away, Away-Home and Away-Draw. In the accuracy formula below, H indicates Home, D indicates Draw, A indicates Away. In addition, the first letter describes the prediction and the second letter the actual outcome,

$$Accuracy = \frac{HH + DD + AA}{HH + HD + HA + DH + DD + DA + AH + AD + AA} \times 100\%.$$

Subsequently, the average return on investment (ROI) is a financial metric that is widely used to measure the probability of gaining a return from an investment. It is a ratio that compares the gain or loss from an investment relative to its cost. The formula of average return on investment is given below,

$$ROI = \frac{\text{Final Value of Investment} - \text{Initial Value of Investment}}{\text{Initial Value of Investment}} \times 100\%.$$

#### 4.1.1 *HOME-approach*

As mentioned in the previous section 44,1% home victories, 25% draws and 30,9% away victories were achieved in the training set. These results perspicuously demonstrate a home advantage during soccer matches. Therefore, our *HOME-approach* indicates the home team always wins.

#### 4.1.2 *BET-approach*

For each match *BET365* generates betting odds for home team wins (B365H), draw (B365D) and away team wins (B365A). According to the bookmaker the event with the lowest odd is the most probable outcome. Therefore, our *BET-approach* indicates the team with the lowest betting odds always wins.

#### 4.1.3 *538-approach*

Our input features consist of *FiveThirtyEight* predictions. *538 home*, *538 draw*, *538 away* indicate the probability home team wins, draw or away team wins respectively. According to *FiveThirtyEight* the event with the highest probability is the most probable outcome. Therefore, our *538-approach* indicates the team with the highest *FiveThirtyEight* probability always wins.

#### 4.1.4 Random forest

Based on the best performing model of Stübinger et al. (2020) we include the random forest in our study. The random forest, created by Breiman (2001), is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently, and with the same distribution for all trees in the forest. This collection of tree predictors $h(x; \theta_k)$ with $k = 1, ..., K$ represents the observed input vector of length $p$ with associated random vector $X$ and the $\theta_k$ are independent and identically distributed random vectors. As mentioned, we focus on the regression setting for which we have a numerical outcome $Y$. The observed data is assumed to be independently drawn from the joint distribution of $(X, Y)$ and comprises $n(p + 1)$-tuples

$(x_1, y_1), ..., (x_n, y_n)$. Eventually, the random forest regression prediction is the unweighted average over the collection:

$$h(\bar{x}) = (1/K) \sum_{k=1}^{K} h(x; \theta_k).$$

The advantage of this technique is that for a large number of trees, it follows from the Law of Large Numbers that random forests do not overfit as more trees are added, but produce a limiting value of the generalization error (Hall and Holmes, 2003). We applied the 'ranger' function in R (Wright, Wager and Probst, 2019), where we implement a grid search for the following parameters: the number of randomly selected predictors, *mtry* $\in \{1, 2, 3, 4, 5, 6, 7\}$. In addition, the *splitting rule* is set to 'variance' and the *minimal node size* is set to 5 for regression.

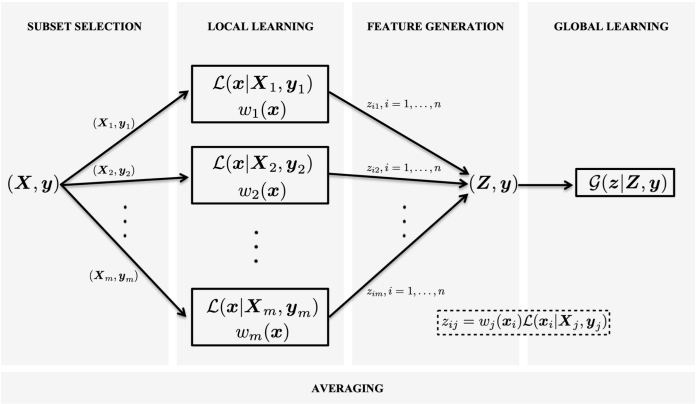
### 4.1.5 Stochastic gradient boosting

Based on the best performing model of Geurkink et al. (2021) we include the stochastic gradient boosting in our study. The stochastic gradient boosting, developed by Friedman (2002), constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current 'pseudo'-residuals by least-squares at each iteration. The pseudo-residuals are the gradient of the loss functional being minimized, with respect to the model values at each training data point, evaluated at the current step. Whereas random forests build an ensemble of deep independent trees, stochastic gradient boosting build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. We applied the 'gbm' function in R (Ridgeway, 2004), where we implement a grid search for the following parameters: the maximum depth of variable interactions, *interaction.depth* $\in \{1, 2, 3, 4\}$ and the total number of trees to fit, *n.trees* $\in \{50, 100, 150, ..., 1500\}$. In addition, *shrinkage* is set to 0.1 and *n.minobsinnode* is set to 20.

### 4.1.6 LESS model

Birbil et al. (2021) developed a two-layer meta-learning algorithm that learns from a set of input-output pairs, where the relation between the input variables and the output variable exhibits a heterogeneous behavior across the predictor space. As this is the case in soccer, we apply this new algorithm in our study to predict the outcome of soccer matches. Figure 4 shows a schematic overview of the LESS model, where a subdivision is made in the following five steps: subset selection, local learning, feature generation, global training, and averaging.

Figure 4: Depiction of the LESS model


SUBSET SELECTION
LOCAL LEARNING
FEATURE GENERATION
GLOBAL LEARNING
$(\boldsymbol{X}, \boldsymbol{y})$
$(\boldsymbol{X}_1, \boldsymbol{y}_1)$
$(\boldsymbol{X}_2, \boldsymbol{y}_2)$
$(\boldsymbol{X}_m, \boldsymbol{y}_m)$
$\mathcal{L}(\boldsymbol{x}|\boldsymbol{X}_1, \boldsymbol{y}_1)$ $w_1(\boldsymbol{x})$
$\mathcal{L}(\boldsymbol{x}|\boldsymbol{X}_2, \boldsymbol{y}_2)$ $w_2(\boldsymbol{x})$
$\mathcal{L}(\boldsymbol{x}|\boldsymbol{X}_m, \boldsymbol{y}_m)$ $w_m(\boldsymbol{x})$
$z_{i1}, i = 1, \ldots, n$
$z_{i2}, i = 1, \ldots, n$
$z_{im}, i = 1, \ldots, n$
$(\boldsymbol{Z}, \boldsymbol{y})$
$\mathcal{G}(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{y})$
$z_{ij} = w_j(\boldsymbol{x}_i)\mathcal{L}(\boldsymbol{x}_i|\boldsymbol{X}_j, \boldsymbol{y}_j)$
AVERAGING

*Notes.* A schematic overview of the LESS model is shown from Birbil et al. (2021), where a subdivision is made in the following five steps: subset selection, local learning, feature generation, global training, and averaging.

Foremost, the subset selection step generates subsets that are concentrated around random points in the input space $(X_j,\, y_j) \subseteq$ (X, y) for $j = 1, ..., m$ by selecting the $k$-nearest samples. Afterwards, a local predictor is trained over each sample subset (denoted by $\mathcal{L}(x|X_m, y_m)$) and transformed into a modified feature space of weighted predictions $w_j(x)$. This weighting is performed such that the local predictions for that input are assigned importance according to the distances of their corresponding subsets from the input point. Then, for each input $x_i$ a feature vector of weighted predictions is generated as $z(x_i) = [w_1(x_i)\mathcal{L}(x_i|X_1, y_1), ..., w_m(x_i)\mathcal{L}(x_i|X_m, y_m)]$, and is written more simply as follows $z_{ij} = w_j(x_i)\mathcal{L}(x_i|X_j, y_j)$ for $j = 1, ..., m$. Subsequently, at the global level an aggregate prediction is performed from the generated feature vector, $\mathcal{G}(z|Z, y)$. Due to the fact that the subsets are chosen randomly, the LESS model benefits from averaging over repetitions to reduce the variance, and then the overall prediction with averaging leads to $\dfrac{1}{r}\, \sum_{l=1}^{r}\, \mathcal{G}(z^{(l)}|Z^{(l)}, y)$, where $r$ corresponds to the number of repetitions.

The advantage of the LESS model is that is it efficient in terms of computation time and it allows a straightforward parallel implementation. We applied the 'LESSRegressor' function in Python (Birbil, 2021) with the following parameter: 633 neighbors ($k$), 19 subsets ($m$) and 20 replications ($r$). The weights are chosen as follows:

$$w_j(x) = \frac{e^{-\lambda d(x, \bar{x}_j)}}{\sum_{j'=1}^{m} e^{-\lambda d(x, \bar{x}_{j'})}},$$

where the distance function $d(x, x') = ||$x - x'$||_2$ along with $\lambda = m^{-2}$.

## 4.2 Tradings strategy

Besides predicting the outcome of soccer matches, we consider two applications to achieve a positive return on investment with betting. Our first trading strategy places the same amount of money on each match. In addition, we generate a second strategy where we double the amount of investment if the match prediction outcome $y$ is located in the upper 10% of the home and away boundaries according to the in-sample forecast with our trainset. So we only double the amount of investment for some home and away predictions, and omit draws. Table 5 shows the amount of investments of the second trading strategy where 2* indicates that the amount of investment is twice the value of 1*.

Table 5: Amount of investments of the second trading strategy

| | AWAY (2*) | AWAY (1*) | DRAW (1*) | HOME (1*) | HOME (2*) |
|---|---|---|---|---|---|
| RF | $\langle -\infty, -1.921 \rangle$ | [-1.921, -0.236] | $\langle -0.236, 0.433 \rangle$ | [0.433, 2.360] | $\langle 2.360, \infty \rangle$ |
| GBM | $\langle -\infty, -1.346 \rangle$ | [-1.346, -0.055] | $\langle -0.055, 0.395 \rangle$ | [0.395, 1.771] | $\langle 1.771, \infty \rangle$ |
| LESS | $\langle -\infty, -1.376 \rangle$ | [-1.376, -0.023] | $\langle -0.023, 0.383 \rangle$ | [0.383, 1.823] | $\langle 1.823, \infty \rangle$ |

*Notes.* The amount of investments of the second trading strategy for the random forest, the stochastic gradient and the LESS model are shown. 2* indicates that the invested amount is twice the value of 1*.

## 4.3 Bootstrapping statistics

Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples, using replacement during the sampling process. We include bootstrapping in our study to find out how reliable and valid the accuracy and return on investments performance metrics are (Freedman, 1981). We replicate 9999 resampled datasets of the outcomes of 1405 soccer matches, and obtain a 95% confidence interval of our estimated accuracies and return on investments through the ‘nptest: Nonparametric Bootstrap and Permutation Tests’ package from Helwig (2021). In addition, we apply a Shapiro-Wilk normality test (Shapiro and Wilk, 1965) and a Q-Q plot (Wilk and Gnanadesikan, 1968) to check for normality for all approaches/models.

# 5 Results

In this section, we conduct a formal investigation to determine which approach/model is the best at predicting the outcome of soccer matches by comparing the accuracy, root mean squared error and median absolute deviation of six approaches/models described in the previous section: *HOME-approach*, *BET-approach*, *538-approach*, random forest, stochastic gradient boosting and the LESS model. Furthermore, we examine the return on investments based on our trading strategies, analyze our bootstrapping results and investigate the models' variable importance.

## 5.1 Soccer match outcome predictions

Table 6 shows the results of the applied approaches/models to predict the outcome of soccer matches. The best obtained parameters are shown in parentheses, where *n.t* is the abbreviation of *n.trees* and *i.d* of *interaction.depth*. Based on the accuracy the stochastic gradient boosting performs best with 53,67%, followed by the LESS model and the *BET-approach* with an accuracy of 53,17% and 52,95% respectively. The *HOME-approach* performs the worst with an accuracy of 44,06%, which was expected because of the naïve assumption a home team always wins.

Figures 10 to 15 in Appendix A show the distribution of the accuracy for the 6 approaches/models through bootstrapping, where we create a 95% confidence interval with all 9999 repeated accuracy outcomes. Appendix B shows the Shapiro-Wilk normality test outcomes and Q-Q plots from which we conclude that each bootstapping is normally distributed. Eventually, we find out that the stochastic gradient boosting also achieves the highest accuracy of 95% confidence interval compared to the other approaches/models: *Home-approach* [41,42%; 46,69%], *Bet-approach* [50,25%; 55,52%], *538-approach* [49,18%; 54,38%], random forest [47,47%; 52,67%], stochastic gradient boosting [50,96%; 56,16%] and the LESS model [50,46%; 55,66%]. Subsequently, the stochastic gradient boosting achieves the lowest root mean squared error of 1.627, followed by the LESS model and the random forest with 1.636 and 1.645 respectively. Although the stochastic gradient boosting has a higher accuracy and a lower root mean squared error than the LESS model, the LESS model obtains the lowest median absolute deviation of 1.486. However, this difference is small since it appears only after three decimal places.

Table 6: Soccer match outcome prediction results

| | HOME | BET | 538 | RF$^{(mtry=1)}$ | GBM$^{(n.t=50,i.d=2)}$ | LESS |
|---|---|---|---|---|---|---|
| Accuracy | 44,06% | 52,95% | 51,81% | 50,18% | **53,67%** | 53,17% |
| RMSE | - | - | - | 1.645 | **1.627** | 1.636 |
| MAD | - | - | - | 1.532 | 1.492 | **1.486** |
| ROI (1) | -10,69% | -4,34% | -4,68% | -3,29% | **1,49%** | 0,58% |
| ROI (2) | - | - | - | -3,42% | **1,35%** | 0,34% |

*Notes.* The soccer match outcome prediction results of the *HOME-approach*, *BET-approach*, *538-approach*, random forest, stochastic gradient boosting and the LESS model are shown with best obtained parameters in parentheses. We compared the performance of the model through the accuracy, the root mean squared error (RMSE), the median absolute deviation (MAD) and two trading strategies, ROI (1) and ROI (2).

In spite of the fact that the *BET-approach* has a higher accuracy than the random forest, the random forest achieves a less negative return on investment for both trading strategies, -3,29% for ROI (1) and -3,42% for ROI (2). Moreover, the stochastic gradient boosting and the LESS model are even capable to obtain an average positive return of respectively 1,49% and 0,58% using our first trading strategy where we invest the same amount of money each match. It is worth noting that our second trading strategy, where we double our investments if our model presents more certainty, achieves a lower average return compared to betting the same amount of money on each match. Figures 24 and 25 in Appendix C show the distribution of the return on investments through bootstrapping, where we invest the same amount of money each match for the two model that generate a positive return. Subsequently, we find statistical evidence that both models obtain a 95% confidence interval for the average return on investment of respectively [-4,17%; 7,13%] and [-5,03%; 6,25%] through bootstrapping. Moreover, we find statistical evidence that both models obtain a positive return on investments in respectively 71,2% and 58,1% of the cases. Therefore, we conclude that the stochastic gradient boosting is the best model, according to our performance metrics, for predicting the outcome for soccer matches and is able to obtain a positive return on investment through betting with 71.2% certainty.

Tables 7 and 8 show the confusion matrices of the two best performing models: stochastic gradient boosting and the LESS model. The columns indicate the predicted outcomes and the rows specify the actual outcomes. In both cases the models predict the same amount of correct home team wins. In addition, stochastic gradient boosting predict more draws and away's correct. We include a plot of the outcomes of the first 250 matches in our testset in Appendix D for a visual overview.

Table 7: Confusion matrix of GBM

| | A | D | H |
|---|---|---|---|
| A | **236** | 100 | 80 |
| D | 99 | **107** | 128 |
| H | 100 | 144 | **411** |

Table 8: Confusion matrix of LESS

| | A | D | H |
|---|---|---|---|
| A | **231** | 100 | 82 |
| D | 102 | **105** | 126 |
| H | 102 | 146 | **411** |

Figures 5 and 6 show the distribution of the predicted goal difference between the home team and away team. We found out that the LESS histogram is more symmetrical and looks more like the normal distribution than the stochastic gradient boosting histogram. In addition, in both cases the number of positive values have the largest proportion, and therefore the distribution is shifted to the right. As previously mentioned in the trading strategy section, the boundaries of draw are further to the positive side, so this outcome was expected.

Figure 5: Predicted y distribution from GBM

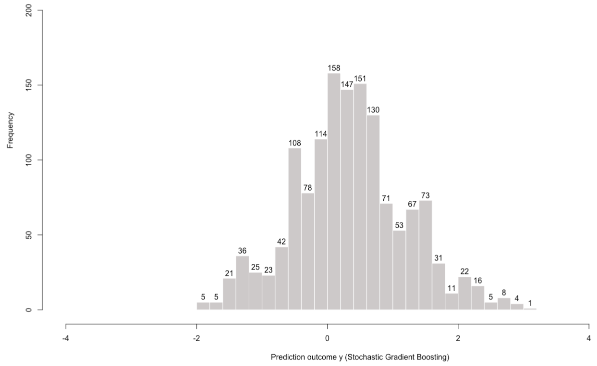
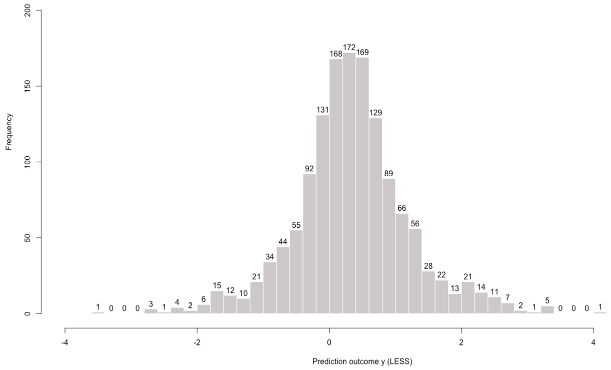


Figure 6: Predicted y distribution from LESS

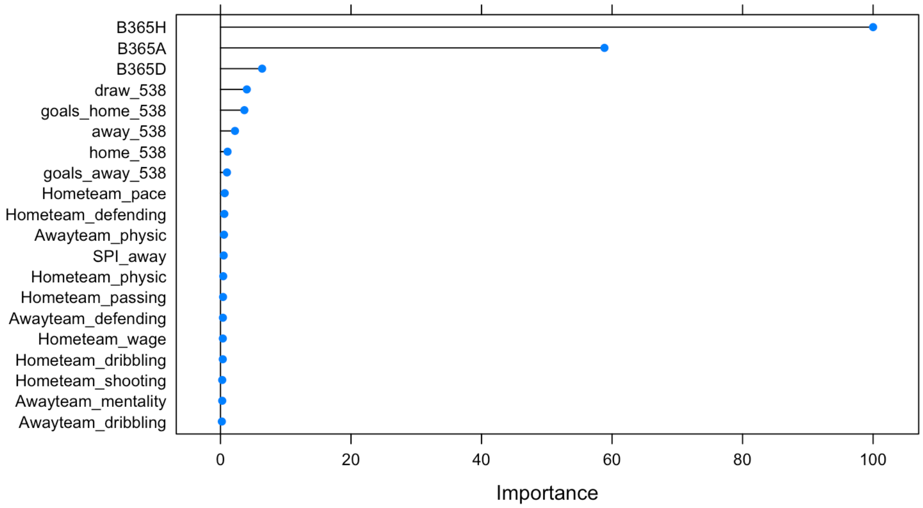


*Notes.* The distribution of the predicted goal difference between the home team and away team are shown for the stochastic gradient boosting and the LESS model.

## 5.2 Variable importance

We apply the two best performing regression models, the stochastic gradient boosting and the LESS model, to determine which characteristics influence the difference in goals scored by the home team and away team most. In Figure 7, we visualize the results of the twenty most influential variables of the stochastic gradient boosting in an importance bar. The most influential input feature is the B365H (100%), followed by B365A (58.84%) and B365D (6,40%). It appears that the three most important characteristics are betting odds input features, which was expected since *BET-approach* achieved the third highest accuracy.

Figure 7: Variable importance of GBM

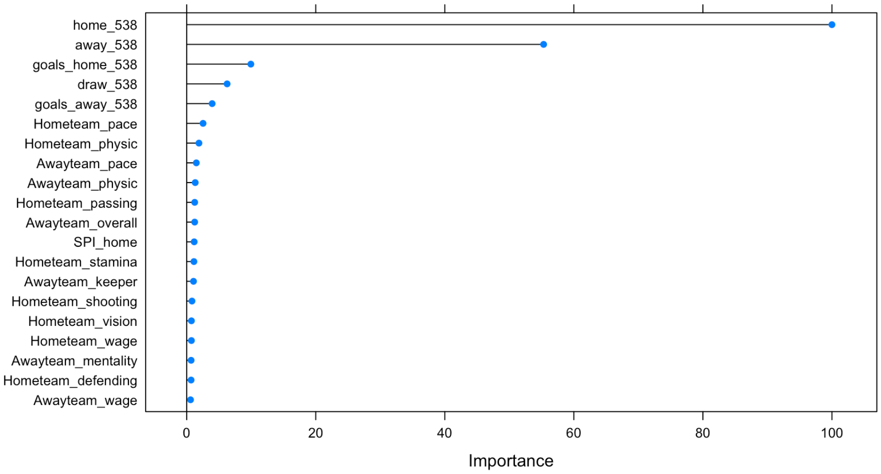


*Notes.* The twenty most influential variables of the stochastic gradient boosting are shown in an importance bar.

Since the most influential stochastic gradient boosting variables are BET365 betting odds, we exclude them in our consecutive model to find out how this adjustment affects the model performance and variable importance. Eventually, the stochastic gradient boosting with only FIFA player characteristics and FiveThirtyEight ratings obtains a root mean squared error of 1,648 and the characteristics that influence the difference in goals scored by the home team and away team most are shown in Figure 8. Here, the 538 home (100%) and 538 away (55,32%) are most influential, followed by three other FiveThirthyEight ratings. Moreover, it is worth mentioning that the stochastic gradient boosting without betting odds performs worse than the random forest, stochastic gradient boosting and the LESS model, with original input features.

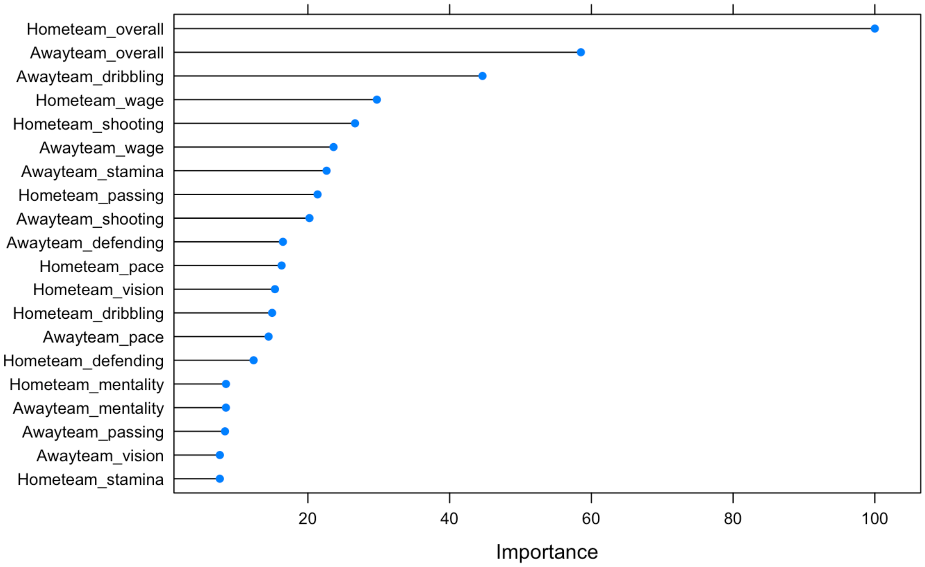
Finally, we construct the stochastic gradient boosting without betting odds and FiveThirtyEight ratings to find out how this adjustment affects the model performance and variable importance. Figure 9 shows the variable importance of this model, where we obtain a root mean squared error of 1,670. As expected the stochastic gradient boosting without betting odds and FiveThirthyEight ratings performs worse than the random forest, stochastic gradient boosting and the LESS model, with original input features.

Figure 8: Variable importance of GBM without betting odds



*Notes.* The twenty most influential variables of the stochastic gradient boosting without betting odds are shown in an importance bar.

Figure 9: Variable importance of GBM without betting odds and 538 ratings



*Notes.* The twenty most influential variables of the stochastic gradient boosting without betting odds and FiveThirthyEight ratings are shown in an importance bar.

Compared to the stochastic gradient boosting the results of the LESS model are easier to interpret due to its local regressions. Since we include 20 replications and 19 subsets, in total 380 local estimators are obtained. Subsequently, for each replication a global estimator is generated and the average is taken. Table 9 shows an example of a local regression for the first replication and the first subsets. As expected, all home team player characteristics have a positive effect on $y$ (the difference of goals between the home team and away team), because a stronger home team has a higher chance of winning. The three most influential home team player characteristics are Overall, Pace and Keeper with subsequently 0.719, 0.725 and 0.796. Moreover, almost all away team player characteristics have a negative effect, with the exception of Pace and Mentality. Also these results are predictably since a stronger away team has a higher chance of winning. Furthermore, the *FivethirthyEight* rating and *Bet365* betting odds results were foreseeable as well. The *SPI home*, *538 home* and *538 goals home* have a positive influence, while all variable with preference for away team wins or draws have a negative effect. In addition, the counterplay between the probabilities for the three outcomes and the resulting bets explains that a increase in B365H results in less chance of winning, and the opposite effect for away and draw.

Table 9: LESS local regression results

| | | Coefficient |
|---|---|---|
| *Fifa player characteristics* | | |
| *Hometeam* | | |
| | Overall | 0.719 |
| | Pace | 0.725 |
| | Skills | 0.324 |
| | Shooting | 0.599 |
| | Passing | 0.647 |
| | Dribbling | 0.714 |
| | Defending | 0.507 |
| | Physic | 0.459 |
| | Wage | 0.155 |
| | Vision | 0.572 |
| | Mentality | 0.391 |
| | Stamina | 0.538 |
| | Keeper | 0.796 |
| *Awayteam* | | |
| | Overall | -0.086 |
| | Pace | 0.134 |
| | Skill | -0.553 |
| | Shooting | -0.148 |
| | Passing | -0.176 |
| | Dribbling | -0.126 |
| | Defending | -0.151 |
| | Physic | -0.008 |
| | Wage | -0.280 |
| | Vision | -0.247 |
| | Mentality | 0.098 |
| | Stamina | -0.179 |
| | Keeper | -0.158 |
| *FivethirthyEight Rating* | | |
| | SPI home | 0.722 |
| | SPI away | -0.006 |
| | 538 home | 0.845 |
| | 538 away | -0.834 |
| | 538 tie | -0.300 |
| | 538 goals home | 0.648 |
| | 538 goals away | -0.678 |
| *Bet365 betting odds* | | |
| | B365H | -0.597 |
| | B365D | 0.183 |
| | B365A | 0.490 |

# 6 Conclusion and Discussion

## 6.1 Conclusion

In this study, we predict the outcome of soccer matches based on FIFA player characteristics, *FiveThirtyEight* team ratings and *Bet365* betting odds. The applied dataset consists of 14.062 soccer matches from eight European soccer leagues from the period July 2016 to March 2022. For evaluation purposes, we randomly split the data into two subsets. The training set consists of 90% of the data, 12.657 matches, and the test set contains the remaining 10%, 1.405 matches.

We apply three naïve approaches and three machine learning models: the *HOME-approach* indicates the home team always wins, the *BET-approach* indicates the team with the lowest betting odds always wins, the *FiveThirtyEight-approach* indicates the team with the highest team rating always wins, random forest which is the best performing model according to Stübinger et al. (2020), stochastic gradient boosting which is the best performing model according to Geurkink et al. (2021), and LESS model from Birbil et al. (2021) which is a machine learning algorithm designed for populations where the relation between the input variables and the output variable exhibits a heterogeneous behavior across the predictor space. The algorithm starts with generating random subsets followed by training a local predictor for each subset. Then those predictors are combined to yield an overall predictor.

To evaluate the performance of the different models we apply four metrics: the root mean squared error; the median absolute deviation; the accuracy; and the average return on investment. The first two metrics obtain the model performance based on the different number of goals scored by the home team and away team. In addition, the accuracy is the percentage of correctly predicted matches outcomes, where we classify (home team wins; draw; away team wins) the outcomes based on the different number of goals scored by the home team and away team, and our boundaries. Furthermore, the average return on investment is a ratio that compares the gain or loss from an investment relative to its cost. We consider two applications to achieve a positive return on investment with betting. Based on the outcome predictions of our applied models, we implement two trading strategies to obtain a positive return on investment: places the same amount of money on each match; and increase the amount of investment if the match prediction outcome probability increases.

We conclude that the stochastic gradient boosting predicts the soccer match outcomes best, followed by the Learning with Subset Stacking model, and that both models are capable to obtain an average positive return on investment of respectively 1,49% and 0,58% per match by using our first trading strategy where we invest the same amount of money each match. Moreover, we find statistical evidence that both models obtain a 95% confidence interval for the average return on investment of respectively [-4,17%; 7,13%] and [-5,03%; 6,25%]. In addition, we obtain statistical evidence that both models achieve a positive investment return in respectively 71.2% and 58.1% of the cases. Therefore, we conclude that the stochastic gradient boosting is the best model for predicting the outcome for soccer matches and is able to obtain a positive return on investment through betting.

## 6.2 Discussion and outlook

To improve our study, we propose further research for the limitations we faced. First, the applied FIFA player characteristics are average scores from the original (the line-up preferred by the coach) players for each club each season. However, it is irrational to assume that a team plays with the same eleven players the entire season due to, for example injuries, personal circumstances or fitness. Therefore, we suggest to apply the average scores of players who actually play during a specific match to make the study more realistic. Second, our accuracy predictions are based on the boundaries, which are obtained by using an in-sample forecast of the whole trainset. However, if we create the boundaries per league or per team, we could get completely different boundaries, and perhaps a higher accuracy and return on investment. Lastly, we include the LESS model of Birbil et al. (2021) where we applied the default LESS linear regression for both local and global estimators with standard weights. However, we did not focus on optimizing these weights by tuning through a grid search. Therefore, we recommend future research to concentrate on optimizing the weights to further improve the results.

# References

Arntzen, H., & Hvattum, L. M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, *21*(5), 449–470.

Birbil, I. (2021). *Less: Learning with subset stacking.* Retrieved from `https://github.com/sibirbil/LESS/`

Birbil, S. I., Yildirim, S., Gokalp, K., & Akyuz, H. (2021). Learning with subset stacking. *arXiv preprint arXiv:2112.06251*.

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

Buursma, D. (2011). Predicting sports events from past results towards effective betting on football matches. In *Conference paper, presented at 14th twente student conference on it, twente, holland* (Vol. 21).

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of english football league matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*(2), 157–168.

Danisik, N., Lacko, P., & Farkas, M. (2018). Football match prediction using players attributes. In *2018 world symposium on digital intelligence for systems and machines (disa)* (pp. 201–206).

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.

Elo, A. E. (1978). *The rating of chessplayers, past and present.* BT Batsford Limited.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, *9*(6), 1218–1228.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Geurkink, Y., Boone, J., Verstockt, S., & Bourgois, J. G. (2021). Machine learning-based identification of the strongest predictive variables of winning and losing in belgian professional soccer. *Applied Sciences*, *11*(5), 2378.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering*, *15*(6), 1437–1447.

Helwig, N. (2021). nptest: Nonparametric bootstrap and permutation tests.

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review.

*Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1380.

Hubáček, O., Šourek, G., & železnỳ, F. (2022). Forty years of score-based soccer match outcome prediction: an experimental review. *IMA Journal of Management Mathematics*, *33*(1), 1–18.

Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 proceedings of the 34th international convention mipro* (pp. 1623–1627).

Hvattum, L. M., & Arntzen, H. (2010). Using elo ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470.

Igiri, C. P., & Nwachukwu, E. O. (2014). An improved prediction system for football a match result. *IOSR Journal of Engineering (IOSRJEN)*, *4*(12), 12–20.

Koopman, S. J., & Lit, R. (2019). Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting*, *35*(2), 797–809.

Kuhn, M. (2009). The caret package. *Journal of Statistical Software*, *28*(5), 17.

Ley, C., Wiele, T. V. d., & Eetvelde, H. V. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, *19*(1), 55–73.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118.

McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction. In *Fifth international conference on information technology: New generations (itng 2008)* (pp. 1194–1197).

Odachowski, K., & Grekow, J. (2012). Using bookmaker odds to predict the final result of football matches. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 196–205).

Owen, A. (2011). Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, *22*(2), 99–113.

Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football result prediction with bayesian network in spanish league-barcelona team. *International Journal of Computer Theory and Engineering*, *5*(5), 812.

Prasetio, D., et al. (2016). Predicting football match results with logistic regression. In *2016 international conference on advanced informatics: Concepts, theory and application (icaicta)* (pp. 1–5).

Reep, C., & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, *131*(4), 581–585.

Ridgeway, G., & Ridgeway, M. G. (2004). The gbm package. *R Foundation for Statistical Computing, Vienna, Austria*, *5*(3).

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 399–418.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Shin, J., & Gasparyan, R. (2014). A novel way to soccer match prediction. *Stanford University: Department of Computer Science*.

Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, *10*(1), 46.

Tax, N., & Joustra, Y. (2015). Predicting the dutch football competition using public data: A machine learning approach. *Transactions on knowledge and data engineering*, *10*(10), 1–13.

Teli, L. K., Zaveri, N., Shinde, P., et al. (2018). Prediction of football match score and decision making process. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(2), 162–165.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis for the analysis of data. *Biometrika*, *55*(1), 1–17.

Wright, M. N., Wager, S., Probst, P., & Wright, M. M. N. (2019). Package 'ranger'. *Version 0.11*, *2*.

Wunderlich, F., & Memmert, D. (2018). The betting odds rating system: Using soccer forecasts to forecast soccer. *PloS one*, *13*(6), e0198668.

Wunderlich, F., & Memmert, D. (2021). Forecasting the outcomes of sports events: A review. *European Journal of Sport Science*, *21*(7), 944–957.

# Appendix A

Figures 10 to 15 show the distribution of the accuracy for the 6 approaches/models through bootstrapping for all 9999 repeated outcomes.

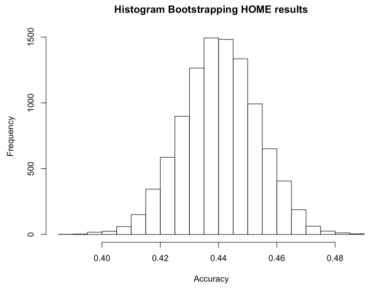


Figure 10: Bootstrapping HOME accuracies

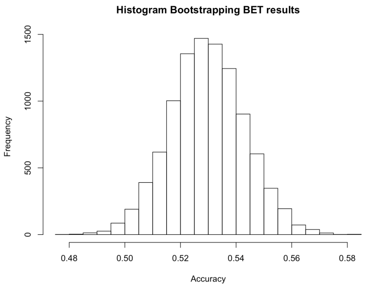


Figure 11: Bootstrapping BET accuracies

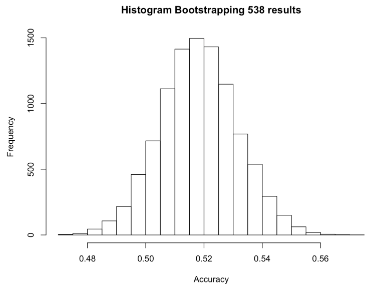


Figure 12: Bootstrapping 538 accuracies

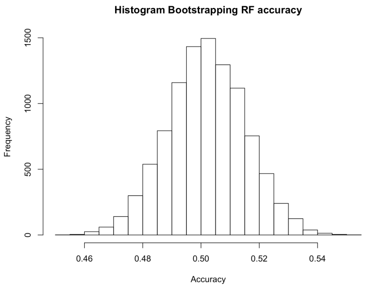


Figure 13: Bootstrapping RF accuracies

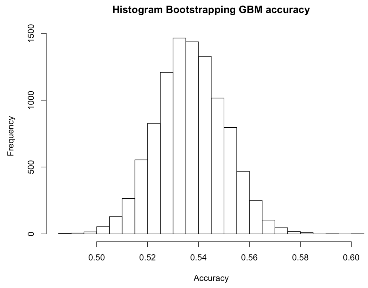


Figure 14: Bootstrapping GBM accuracies

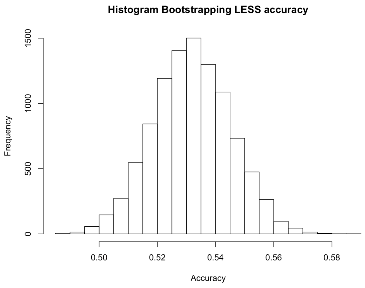


Figure 15: Bootstrapping LESS accuracies

# Appendix B

We apply the Shapiro-Wilk normality test (Shapiro and Wilk, 1965) and the Q-Q plot (Wilk and Gnanadesikan, 1968) for each bootstapping result to check for normality. The Shapiro-Wilk p-value is in all bootstappings greater than 0.05. In addition, we include the Q-Q plot for each approach/model where the linearity of the points suggests that the data are normally distributed. Therefore we assume the data is normal distributed in all cases.

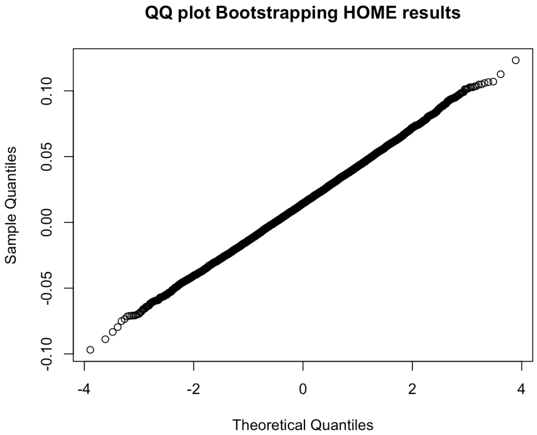


Figure 16: Q-Q plot of the bootstrapping accuracy results of the *HOME-approach*

Shapiro-Wilk normality test of the bootstrapping accuracy results of the *HOME-approach*: p-value = 0.1965

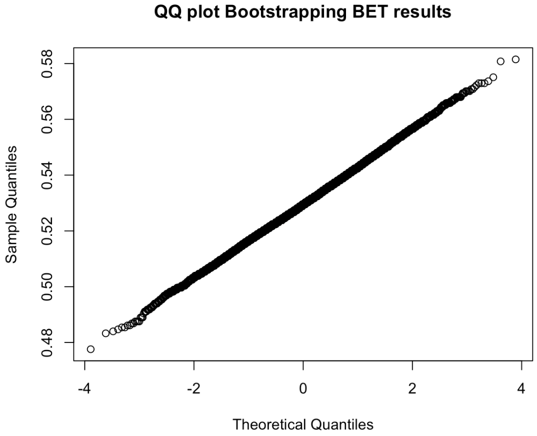


Figure 17: Q-Q plot of the bootstrapping accuracy results of the *BET-approach*

Shapiro-Wilk normality test of the bootstrapping accuracy results of the *BET-approach*: p-value = 0.06826

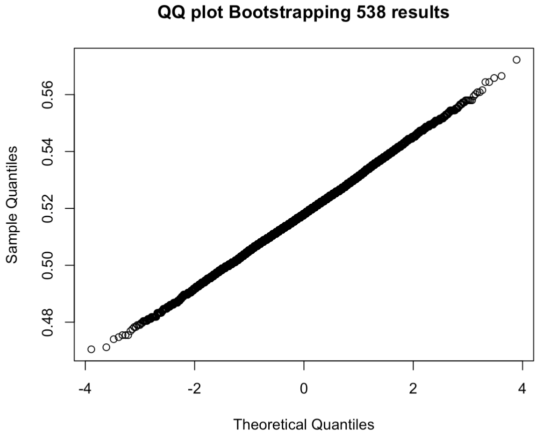


Figure 18: Q-Q plot of the bootstrapping accuracy results of the *538-approach*

Shapiro-Wilk normality test of the bootstrapping accuracy results of the *538-approach*: p-value = 0.05842

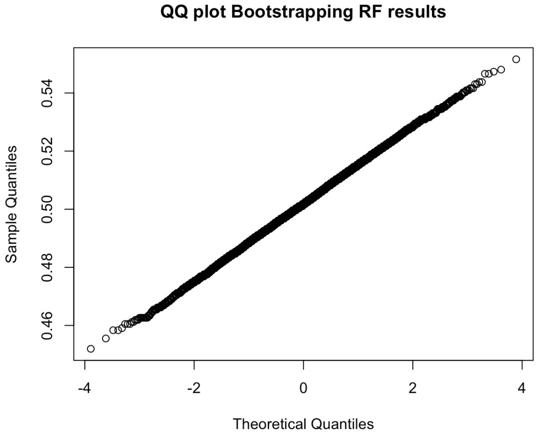


Figure 19: Q-Q plot of the bootstrapping accuracy results of the random forest

Shapiro-Wilk normality test of the bootstrapping accuracy results of the random forest: p-value = 0.1029

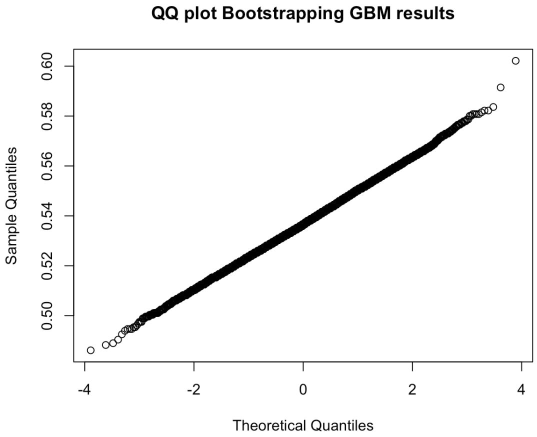


Figure 20: Q-Q plot of the bootstrapping accuracy results of the stochastic gradient boosting

Shapiro-Wilk normality test of the bootstrapping accuracy results of the stochastic gradient boosting: p-value = 0.06987

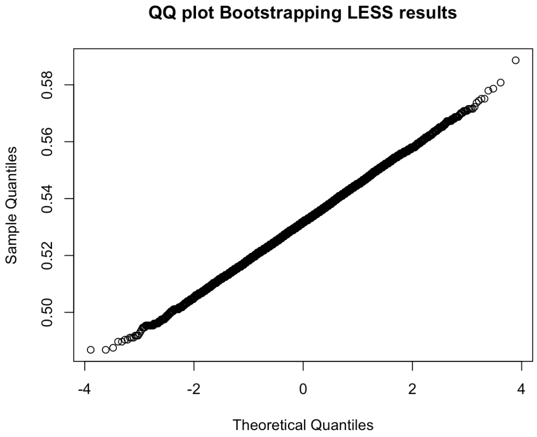


Figure 21: Q-Q plot of the bootstrapping accuracy results of the LESS model

Shapiro-Wilk normality test of the bootstrapping accuracy results of the LESS model: p-value = 0.2584

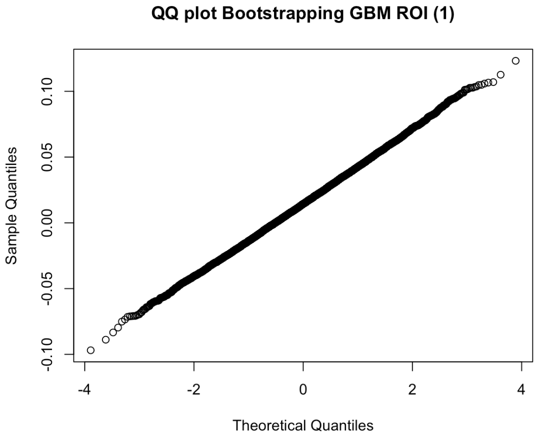


Figure 22: Q-Q plot of the bootstrapping ROI (1) results of the GBM model

Shapiro-Wilk normality test of the bootstrapping ROI (1) results of the GBM model: p-value = 0.07923

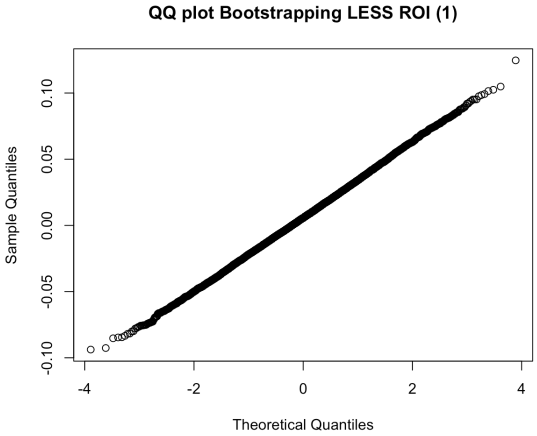


Figure 23: Q-Q plot of the bootstrapping ROI (1) results of the LESS model

Shapiro-Wilk normality test of the bootstrapping ROI (1) results of the LESS model: p-value = 0.2572

# Appendix C

Figures 24 and 25 show the distribution of the return on investment for the 6 approaches/models through bootstrapping for all 9999 repeated outcomes.

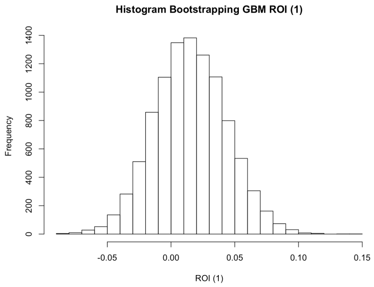


Figure 24: Bootstrapping GBM ROI (1)

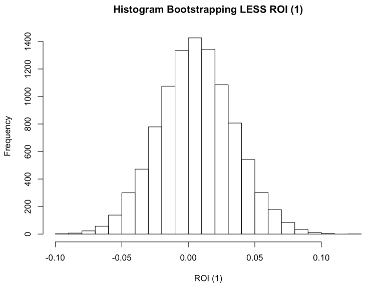


Figure 25: Bootstrapping LESS ROI (1)

# Appendix D

The first 250 predicted soccer matches outcomes of our testset are shown for the random forest, stochastic gradient boosting and LESS model.

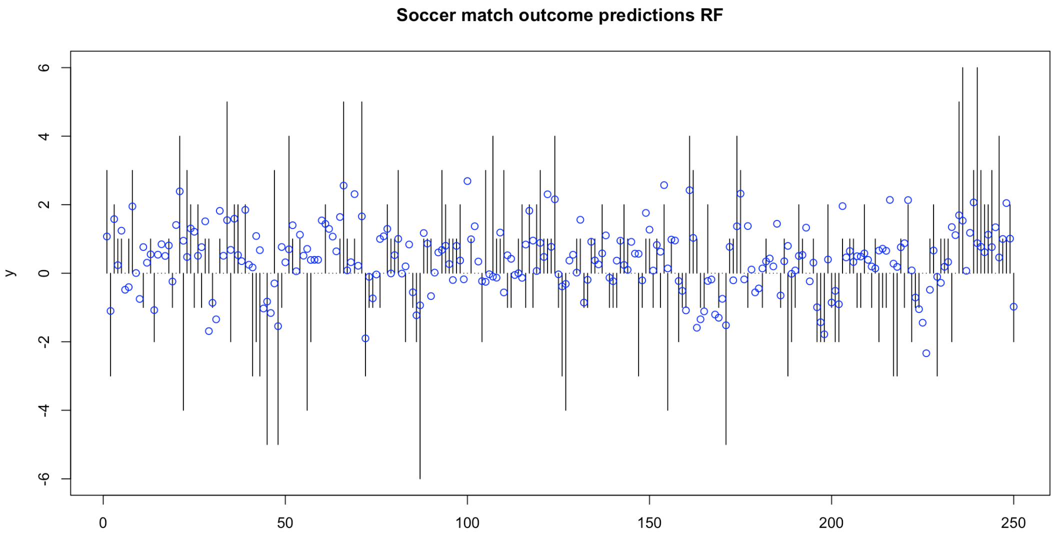


Figure 26: Soccer match outcome predictions RF

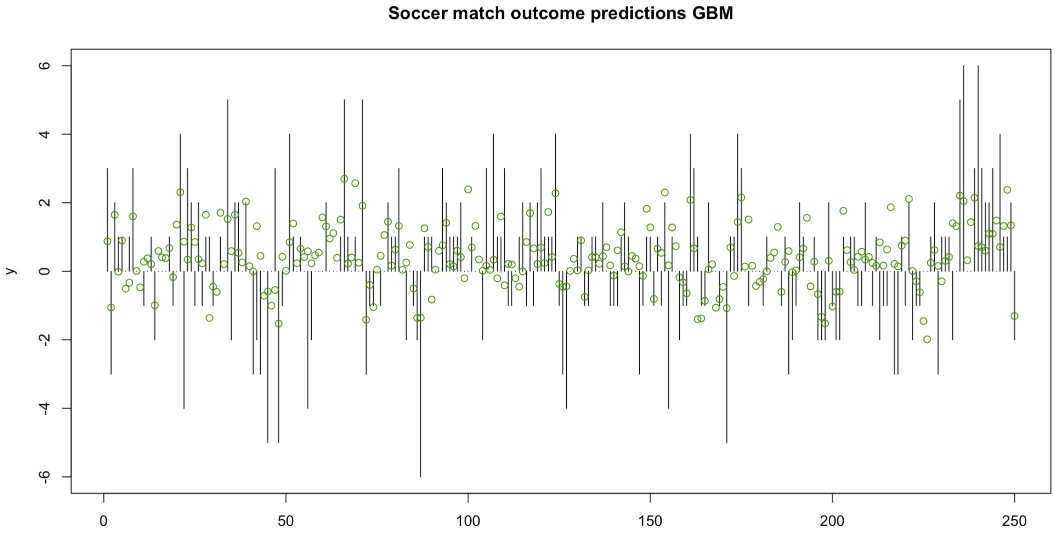


Figure 27: Soccer match outcome predictions GBM

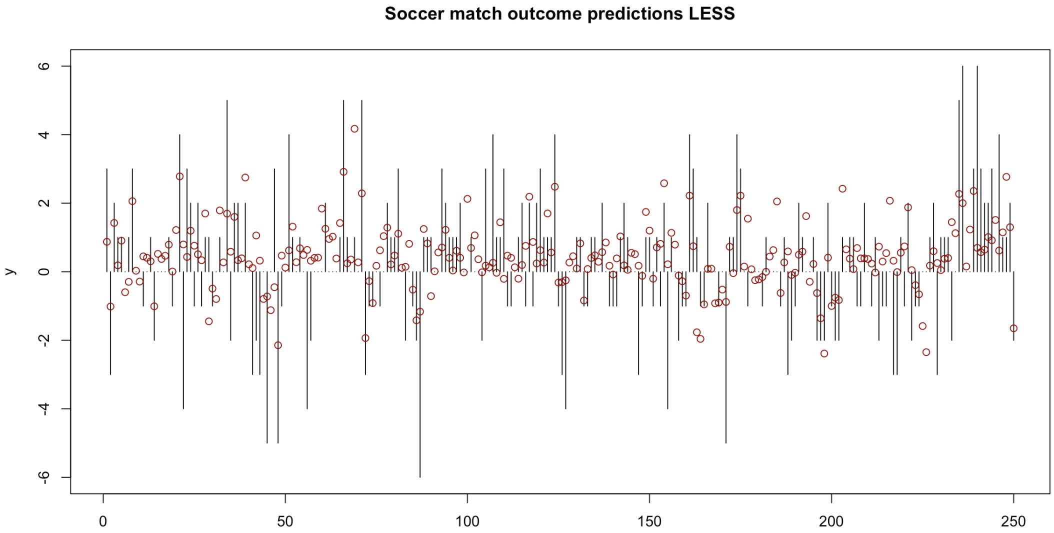


Figure 28: Soccer match outcome predictions LESS