ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Economics & Business

'Beating the Bookmakers using Machine Learning'

Name student:	Dion van Wijk
Student ID number:	477793
Supervisor:	Dr. J.E.M. van Nierop
Second assessor:	Dr. K. Gruber MSc
Date final version:	06/07/2021

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.

Table of Contents

Abstract

In this research paper, the possibility to be profitable in the football betting market is researched. Data of nineteen Premier League seasons ranging from season 2002/2003 until season 2020/2021 have been gathered, which are 7,220 matches in total. Three different machine learning models are used and compared to each other: multinomial logistic regression, random forest, and artificial neural network. Additionally, different betting strategies and money management techniques are compared to each other to see which combination is the most profitable. The last topic which is discussed in this research paper is the most important predictor for the best performing model. The best performing model is the multinomial logistic regression with the fixed betting technique and a betting strategy where you place bets on teams which are mispriced by the bookmakers. This model achieves a return on investment of 28.61% for season 2020/2021 and an average return on investment of around 6% (per annum) over five years. The most important predictors for this introduced model are the head-to-head results. In conclusion, this research paper shows that it is possible to be profitable in the football betting market using machine learning, however there will always be risks involved in the sports betting market.

Keywords: sports betting, bookmakers, machine learning, betting strategies, money management techniques

1. Introduction

The market size of the gambling market is rapidly increasing over the past few years and is expected to grow even further in the future based on the expectations of the European Gaming & Betting Association. The total European total gambling market (regulated market and grey and black markets) was worth 98.6 billion euros in 2019 with online gambling accounting for 24.5 billion euros (European Gaming & Betting Association, 2020). The European market share for the online gambling market is more than half of the global market share, which is about 45.8 billion euros. Moreover, the market share is expected to grow with approximately 68.9% over the next six years (Brandessence Market Research & Consulting Pvt ltd., 2020). The United Kingdom accounted for 30.1% of Europe's gambling market revenue, which is almost three times as much as the second largest contributor Germany (11.4%).

The most popular online gambling activity is sports betting with 41% share of Europe's online revenue in 2019. Sports betting is the activity of placing a bet on the outcome of a particular sports event. Sports betting is hence the most popular online gambling activity, but land-based sports betting (betting at locations which require the physical presence of the player) is still very popular as well. In fact, the global sports betting revenue for online and land-based betting in 2020 was approximately 169 billion euros, which is even more than the Gross Domestic Product (GDP) of Qatar, the organizer of the FIFA World Cup 2022. Speaking of the FIFA World Cup, this event is also the sporting event where the most money is betted. During the final of the World Cup 2018, approximately 6 billion euros was wagered according to a FIFA study (FIFA, 2018). This indicates the amount of money involved in sports betting and especially in (association) football betting.

Bookmakers are the people who facilitate gambling for the bettors and a bookmaker sets odds and pays out winnings on behalf of other people. The bookmakers make money by adjusting the odds as much as possible in such a way that there is an even number of people betting on each team. The overround is the amount by which a bookmakers' odds of a match exceeds the probability of one. The higher the overround, the higher the expected profit for the bookmaker will be (Newall, 2015). In general, the overround is almost always positive which means that there is no opportunity for profitable wagering for the bettor which is confirmed by the research of Sauer (1998). Nevertheless, there are also several research papers

that show that there are arbitrage opportunities using combined betting across different bookmakers (Constantinou & Fenton, 2013; Forrest & Simmons, 2001; Vlastakis, Dotsis, & Markellos, 2006).

In contrast to other financial markets, it is easier to test this market efficiency for the sports betting market because the ex-post realizations are already known after a bet is placed, whereas the ex-post realization is not often known right after an acquisition in other financial markets (Gray & Gray, 1997). This ensures that there is a possibility to test the market efficiency of the sports betting market and see if there are any possibilities to be profitable within this market. Moreover, this research paper will investigate in the different betting strategies to see which strategy is the most profitable.

To accomplish the task of testing the market efficiency of the sports betting market, a model will be introduced to predict match results. Maher (1982) is one of the first researchers that created a model that could predict the scores for football matches. However, in the meantime quite a few other researchers tried to introduce a new or adjusted model for predicting match results. Nowadays, machine learning plays a major role in this field. Several machine learning techniques are used to get the highest accuracy for predicting match results, where the Artificial Neural Network (ANN) is used the most often.

This research paper will use machine learning to research the market efficiency of the sports betting market and see which machine learning technique will produce the highest accuracy. Thereby, the features that are most important for football match prediction will be researched. This research paper will also investigate in the most profitable betting strategy and money management technique. The combination between building a model that can predict match results and researching the best betting strategy using this model will contribute to the existing literature, due the fact that such combinations do not exist already. Moreover, this research paper will be relevant for people who are trying to earn money with sports betting but also for the sports clubs for decision making within the area of tactics. For these reasons, the main research question of this research paper is stated as:

Research question: "To what extent is it possible to be profitable in the football betting market using machine learning?"

The structure of the remaining sections in this research paper are organized as follows. Section 2 will discuss the theoretical framework and academic relevance. The gathering of the data and the used methodology for this research paper is elaborately described in Section 3. Subsequently, Section 4 will formalize the findings based on the performed analyses. Lastly, Section 5 will provide answers to the proposed research question; summarize the research; discuss some limitations; and propose some ideas for further research.

2. Literature Review

This section will discuss the existing literature. The first part of the section will discuss the statistical models; the second part the machine learning models; and the last part will dive into the different money management techniques within sports betting.

2.1. Statistical Models

Maher (1982) is one of the first researchers who made a thoroughly analysed model for prediction of results in football matches. The number of goals scored and conceded by the home team are captured in Poisson distributed variables and Maher (1982) assumes that these variables are independent. The Poisson distribution is a discrete probability distribution where the average time between the events is known, but the exact timing of the events is random. Moreover, the author assumes that each team has an attack and defence strength, where a high attack strength indicates that a team scores many goals, and a low defence strength indicates that a team concedes only a few goals. The last independent variable of the model is the home field advantage, which is assumed to be the same for all teams. He finds that the independent Poisson model gives reasonably accurate description of football scores.

Nonetheless, a fundamental question is whether the underlined distribution can be assumed to be Poisson distributed. Karlis & Ntzoufras (1998) assume that the attack strength of a team is not constant throughout the season, which means that the physical conditions of each time vary over time. This assumption leads to a preference for a mixed Poisson model with a negative Binomial. Another question stated by the authors is whether the number of goals scored by the two opponents in the same match are independent, however they did not find that there is no strong dependence between these two variables. Also, this research found that assuming independent Poisson distributions suffices for match result predictions, however this model had his limitations as well.

Dixon & Coles (1997) introduced a time-dependent model for prediction of results. The authors proposed two important extensions to Maher's model. Firstly, they adjusted the model in such a way that low-scoring draws are slightly more probable and the results 1-0 and 0-1 are slightly less probable. Secondly, they assumed, as Karlis & Ntzoufras (1998), that the attack and defence strength is not constant over time. Additionally, the authors used the bookmakers' odds as independent variables and conclude that the proposed model have a

positive return when using it as the basis of a betting strategy. This indicates that it should be possible to be profitable when predicting football match results.

2.2. Machine Learning Models

Machine learning is part of Artificial Intelligence (AI) and is the study of computer algorithms that improve automatically by the use of data and through experience (Mitchell, 1997). Besides the statistical models, many machine learning techniques have been introduced in the sport result prediction field. All the used machine learning techniques in previous research papers will be discussed briefly in this part of the section. Table 1 shows the theoretical framework of all the existing literature about using machine learning in football result prediction.

The first used machine learning technique for football results prediction is a Bayesian network. Joseph, Fenton, & Neil (2006) tried to predict the results of the English team Tottenham Hotspurs for the period 1995-1997 using a Bayesian network. The authors conclude that the proposed model outperformed other machine learning techniques such as K-Nearest Neighbours (KNN), naive Bayesian learner and MC4 decision trees. The accuracy of the introduced model is 59.21%, however a limitation of this research is that the model is specific for just one team during that specific period.

Owramipur, Eskandarian, & Mozneb (2013) also proposed a Bayesian network for predicting match results of the Spanish team FC Barcelona. The authors added more features to the model like the weather conditions, psychological state of players, and whether or not any of the main players are participating in the match. The accuracy of this model is about 92%, however the period is just one season and also this model is specific for just one team.

Nevertheless, there are also research papers which tried to predict match results for multiple teams. Buursma (2011) used a logistic regression to predict the Dutch football competition and used fifteen years of data. The prediction accuracy of this model is about 55%, which is lower than what the author hoped to generate beforehand, however the research shows that with the right betting strategy the model can lead to profits in the long term.

Prasetio & Harlili (2016) also used a logistic regression to predict football match results. The authors tried to predict the match results of the Premier League for the season 2015/2016 and used data from season 2010/2011 up to and including season 2015/2016. This research paper is different from others as the research uses only significant variables gathered from

research papers in the same field. Using a logistic regression, the authors built a model with a prediction accuracy of 69.5% with the defence strength of the home and away team as significant variables.

The Support Vector Machine (SVM) is used as well to predict the match results of football matches. Igiri (2015) used this machine learning technique to predict fifteen matches in the Premier League. The performance of the SVM proposed in this research paper showed a prediction accuracy of 53.3%, which is relatively low. The author concludes: "Until proven otherwise by other studies, an SVM-based system (as devised here) is not good enough in this application domain."

According to Bunker & Thabtah (2019), the most used machine learning technique in the field of sport results prediction is the Artificial Neural Network (ANN). Rahman (2020) used an ANN to predict the winners of the matches in the group stage of the FIFA World Cup 2018. The author used the FIFA Soccer Rankings and historical international results from 1872 until 2018. The ANN proposed by this research paper resulted in a prediction accuracy of 63.3%, but the author states that this accuracy can be increased with more accurate information of the teams. He also states that machine learning – and especially deep learning – can be used for successfully predicting outcomes of football matches.

Pettersson & Nyquist (2017) used Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) for predicting the outcomes of football matches. The data set which is used in this research is extremely large consisting historical match results of multiple seasons of leagues from 63 different countries. Thereby, the authors compared different approaches of these machine learning techniques and compared these results to naive statistical models and human accuracy. The classification accuracy of this research lies around 50% before the game has started. They also predicted the number of goals scored by both teams which resulted in a lower accuracy when the game has not started yet. However, in this research paper the authors also used the data from during the game which resulted in better predictions (up to 98.63%), since more information about the game is present.

Alfredo & Isa (2019) used a tree-based model algorithm for football match prediction. C5.0, random forest, and extreme gradient boosting are compared to see which algorithm generated the highest accuracy. The training period of this research has a period of ten seasons and the model includes fifteen different features to predict the match results. The best

performing algorithm is the random forest algorithm with an accuracy of 68.55%. The authors conclude that the tree-based algorithms are not good enough in predicting a football match result because at that time there were already research papers with a higher accuracy. These research papers will described in the following paragraphs.

Beside using only one machine learning model, there are also many researchers who combined or compared several machine learning techniques with each other. Igiri & Nwachukwu (2014) used an ANN and a logistic regression, which yielded in accuracies of 85% and 93% respectively and were higher than the existing models. The authors introduced new features to add to the model which have not been used in previous models. Igiri & Nwachukwu (2014) added the players' performance index, managers' index, and bookmaker odds. Adding these features to the model yielded in a higher accuracy, but a limitation of this research is that the availability over a longer period is limited.

Tax & Joustra (2015) compared nine different classification algorithms. The authors used data from thirtheen years of the Dutch football league and were interested in the difference in prediction accuracy between a model with betting odds alone and a hybrid model of betting odds and other match features. The Naive Bayes and ANN were the best performing classifiers with an accuracy of 54.7% on the full features set, which is relatively low compared to other techniques. Both classifiers were combined with a Principal Component Analysis (PCA).

A research paper that also compared many machine learning techniques with each other is the research of Hucaljuk & Rakipovic (2011). The authors compared six different techniques, including Bayesian networks, KNN, random forest, and ANN. In the end, the ANN performed the best with an accuracy of 68%, although also this research has some limitations. The author states that including the form of each player in the match could probably lead to better results and a larger data set would help the model to train better.

Zaveri, Shah, Tiwari, Shinde, & Teli (2018) also compared multiple machine learning techniques. The authors researched the difference in performance for the techniques: logistic regression, random forest, ANN, linear SVM, and naive Bayes. Features as match history, goals history, players stats (from FIFA 18), and team stats (from FIFA 18) are used to predict the winner of football matches in the Spanish league. The aim of this research was to improve the decision-making system for football team managers in the field of team selection, tactic

selection, player evaluation etc. The machine learning technique with the highest prediction accuracy is the logistic regression. The logistic regression achieved an accuracy of 71.63% using all data bases. All previous discussed research papers are shown in a theorical framework in Table 1. This table shows all the existing literature with their used machine learning model, the achieved accuracy, and some limitations. The average of the accuracies is around 65%, however all research papers have their limitation.

Table 1: Theoretical framework of existing literature about using machine learning in football results prediction, including the machine learning technique; the accuracy and the limitations of the research. The table is ordered by the year of publication.

Research	Technique	Accuracy*	Limitations
Joseph, Fenton, & Neil, 2006	Bayesian Network	59.21%	Specific for one team.
Buursma, 2011	Logistic regression	55%	Relatively low accuracy.
Hucaljuk & Rakipovic, 2011	ANN	68%	Relatively little data; room for improvement of the feature selection.
Owramipur, Eskandarian, & Mozneb, 2013	Bayesian Network	92%	Specific for one team; relatively short period.
Igiri & Nwachukwu, 2014	ANN and logistic regression	85%; 92%	Relatively short period.
Igiri, 2015	SVM	53.30%	Relatively low accuracy; relatively short period.
Tax & Joustra, 2015	Naïve Bayes and ANN	54.70%	Relatively low accuracy.
Prasetio & Harlili, 2016	Logistic regression	69.50%	Relatively little predictive features.
Pettersson & Nyquist, 2017	RNN and LSTM	~ 50% - 98.63%	Relatively short period; relatively low pre-game accuracy.
Zaveri, Shah, Tiwari, Shinde, & Teli, 2018	Logistic regression	71.63%	Use of a game database which can differ from real world statistics.
Alfredo & Isa, 2019	Tree-based algorithms	68.55%	Relatively low accuracy; relatively little predictive features.
Rahman, 2020	ANN	63.30%	Specific for international tournaments; room for accuracy improvement.

* The number of decimals is rounded on two digits, except in the cases the accuracy is exactly equal to the value.

This part of the section and Table 1 show that there are already some research papers in predicting football match results. Thereby, all these research papers used different machine learning techniques with some limitations. The best performing techniques in the existing literature are the neural networks and the logistic regression. Moreover, several machine learning techniques can also be combined to get higher prediction accuracy. The first sub-question of this research paper will focus on finding the best performing machine learning technique for predicting football match results and is stated as:

Sub-question 1: "Which machine learning model will perform the best in predicting football match results?"

Besides the different machine learning techniques, all previous research papers used many different features for predicting football matches. The most research papers used betting odds and historical results for the predicting problem. Some of them also included the form of the teams to the model. Nevertheless, there are even more predictors which can be added to the model. Godin, Zuallaert, Vandersmissen, De Neve, & Van de Walle (2014) for example, they used tweets to predict the English Premier League. Besides the statistical analysis, the author used the Twitter volume, sentiment analysis, and user prediction analysis for predicting. The conclusion of this research is that the proposed model can beat the experts and the bookmakers, which would result in a profit of 30%. Kampakis & Adamides (2014) also confirm with their research that there is evidence that Twitter can provide useful information for the prediction of football outcomes.

Aloufi & El Saddik (2018) did also used football-specific tweets to predict match results for the FIFA World Cup 2016 and the Champions League season of 2016/2017. Thereby, different sentiment lexicons were used to create a new football-oriented sentiment lexicon. Combining this lexicon with a SVM led to the highest accuracy compared to a multinomial naive Bayes classifier and a random forest.

As discussed, there are many features – like betting odds, historical results, and sentiment from tweets – which can be used to predict football match results. However, so far there is no research papers which combine all those features and research which feature has the highest contribution in this area. This research paper will elaborate on this and therefore the second sub-question of this research is:

Sub-question 2: "Which predictor(s) has/have the most effect on predicting football match results in the introduced model of this research?"

2.3. Money Management Techniques

There are many betting strategies – also called money management techniques – which can be used to make money with sports betting. Several money management techniques will be discussed in this part of the section.

The most obvious money management technique is called fixed amount betting. When using this strategy, the bettor allocates the same amount to each bet. A similar approach is proportional betting, though with this strategy the bettor does not allocate the same amount

of money to each bet but the same proportion of money to each bet. The approach when you provide each bet of the same expected return is called fixed return betting. All these money management techniques are quite simple, however there are also some more advanced money management techniques.

One of those more advanced techniques is the Fibonacci sequence technique. Each subsequent number in this sequence is the sum of the previous two numbers in the sequence (Sigler, 2002). You start with betting the first number of the sequence on a bet. If you lose, you place the second number of the sequence on a bet and so on. If you win, you move down two units and place that number of the sequence on a bet. In theory, the gambler with an unlimited bankroll will eventually win, because with a single bet all previous lost money can be earned back. However, many bookmakers have a betting limit on the amount they will accept.

The Kelly Criterion is also used often as money management technique. This criterion, proposed by Kelly (Kelly Jr, 2011), is used to determine what proportion should be bet on which bet. The formula to calculate this amount is shown in equation (1), where f is the fraction of the current bankroll to wager; p is the probability of a win; and b is the net fractional odds received on the wager. If $f < 1$, you should not bet anything on the wager, whereas if $f > 1$, you should bet fraction f of your current bankroll on the wager.

$$f = \frac{p(b+1)-1}{b} \tag{1}$$

Another money management technique is the variance-adjusted technique, which is introduced by Rue & Salvesen (2000). This technique is a simplification of the Markowitz portfolio management (Markowitz, 1952) and looks at the difference between the expected profit and the variance of the profit. Rue & Salvesen (2000) wanted to minimize this number to make the most profit. The stake is calculated by equation (2), where c_i is the inlay (stake) on the bet, w_i is the bookmaker odds, and p_i is the probability of winning. The authors conclude in their research paper that single bets are more profitable than combinations bets. For the season 1997/1998 in the Premier League, they would have a profit of 39.6% using this money management technique.

$$c_i = (2w_i \ 1 - p_i \ )^{-1} \tag{2}$$

Langseth (2013) compared several money management techniques with each other for the seasons 2011/2012 and 2012/2013 in the Premier League. The results for both seasons are quite different, in fact for the first season all techniques are profitable, whereas for the second season only the variance adjusted technique is profitable (9.1%). Thereby, the most profitable technique in the first season is the fixed return technique with a profit of 24.2%. As stated in Langseth (2013), the results among the two seasons differ and there is no clear most profitable money management technique. Therefore, the last sub-question of this research paper is about researching the most profitable money management technique using the introduced model and is:

Sub-question 3: "Which money management technique in football betting generates the most profit using the introduced model of this research?"

3. Data & Methodology

This section will dive into the data and methodology part. The first part of the section will focus on the data which is gathered and will discuss some statistics of this data set. The second part is about the methodology used in this research paper.

3.1. Data

The data part is segregated into two parts: data collection and descriptive statistics. The first part will explain how the data set is constructed and where the data is from. The second part will discuss the descriptive statistics of this data set and will give some small insights based on these statistics.

3.1.1. Data Collection

The data which are used for this research is from the Premier League, which is the first league of football in England. Data of nineteen seasons in total ranging from season 2002/2003 up to and including season 2020/2021 have been gathered. The data set contains data of 7,220 Premier League matches in total, with each season having 380 matches. This number of matches is based on the twenty clubs in the Premier League each season. The most important data are gathered from www.football-data.co.uk/data.php. Variables like the date of the matches, playing teams, match results, and the betting odds are gathered from this website. The match statistics of each match are also present in this data set however these are not used for this research because this information is not present prior the start of a match. The full-time betting odds data of the data set are the betting odds of Bet365. Bet365, which is founded in 2002, is a British online gambling company based in the United Kingdom and offers sports betting and casino type games. The betting odds for weekend games are collected Friday afternoons, and for midweek games on Tuesday afternoons. The match round number is added manually to this data set.

Based on the match results of the present data, the ranking of the home and away team can be calculated after each match round. The attack strength and defence weakness are also calculated for both teams. The attack strength is the team's average number of goals, divided by the league's average number of goals. The defence weakness is the team's average number of goals conceded, divided by the league's average number of goals conceded. The

higher the attack strength and the lower the defence weakness, the better these scores are. All these added variables are calculated based on all previous matches of that season. For the first match round of a season there is no data to use to calculate these variables, because not every team has played at least one game at that moment. For those first ten matches, the variables are set to zero.

Moreover, several other variables are calculated in the same way as previous variables. For example, the average number of points and the number of losing points is calculated for the home and away team. The average number of points is calculated by dividing the number of points of a team by the number of matches played by that team. The average number of losing points is calculated by dividing the number of losing points by the number of matches played by that team. The number of losing points is the maximum possible number of points minus the actual number of points achieved by a team. The average goal difference of both teams is also calculated by dividing the goal difference points by the number of matches played by that team. The goal difference is the number of goals scored minus the number of goals conceded. For all those variables the average is taken to have the same scale over the whole season.

Based on the data of that season, the ratio of home wins, home draws, away wins, and away draws are calculated. The ratio of home wins is calculated by the number of home wins divided by the number of home games. The other variables are calculated with the same technique. The data of only the specific season are used to include some way of the form of a team. If you would use all historic data, you may include information which is no longer relevant because of a significant change in the combination of team players in a team or other changes.

Besides these calculated variables, some variables are added to the data set using web scraping. The first variable which is scraped from the web is the match attendance. The match attendance of each match is scraped from https://www.worldfootball.net. However, for some matches there were no spectators allowed (due to COVID-19 measures). The information for the match attendance for these matches on the website is 'without spectators'. All matches with this value are converted to zeros.

The Elo ratings for the home and away team are also scraped from the internet. The Elo rating system is a method for calculating the relative skill levels of players/teams. The

website http://clubelo.com/ENG contains the Elo rating for many clubs per date. The Elo ratings on this website are updated after each match, including international tournaments and national cup games. Based on these Elo ratings, the probability of winning, drawing, or losing a game can be calculated and added to the data set. However, the Elo rating system is originally designed as a method for calculating the relative skill levels of players in zero-sum games such as chess. With zero-sum games, one person gains, and another person loses, which results in a zero-net benefit for both players. Because the Elo rating was designed to analyse the winning percentage of a board game that have rare draw games, the probability of drawing is not specified in the original Elo system. Therefore, Xiong, Yang, Zin, & Iida (2016) proposed new equations to calculate the probabilities. The equations for drawing, winning, and losing are shown in equation (3), (4), (5), respectively. Hvattum & Arntzen (2010) state in their research paper that a home field advantage should be added to the Elo rating of the home team when calculating the probabilities. In their research, they propose a constant value of 80 points to add, however http://clubelo.com/ENG also have data available for the home field advantage per day, so this value is also scraped per day and used when calculating the probabilities, meaning that the home field advantage is not a constant data point but a variable data point. The descriptive statistics of this variable and all other variables will be discussed in the next part of this section.

$$P(\text{draw}) = \frac{1}{\sqrt{2\pi e}} \exp\left(-\frac{\left(\frac{\text{Elo}_{\text{home}} + \text{HomeFieldAdvantage} - \text{Elo}_{\text{away}}}{200}\right)^2}{2e^2} \right) \tag{3}$$

$$P(\text{win}) = \frac{1}{1 + 10^{\frac{\text{Elo}_{\text{away}} - \text{Elo}_{\text{home}} + \text{HomeFieldAdvantage}}{400}}} - \frac{1}{2} P(\text{draw}) \tag{4}$$

$$P(\text{lose}) = \frac{1}{1 + 10^{\frac{\text{Elo}_{\text{home}} + \text{HomeFieldAdvantage} - \text{Elo}_{\text{away}}}{400}}} - \frac{1}{2} P(\text{draw}) \tag{5}$$

A variable that is calculated in the same way as the ratios is the form of a team. The form is indicated by the mean of the number of points achieved in the previous five matches multiplied by the mean of the Elo rating of the opponents for the home playing team. The mean of number of points is multiplied by the mean of the Elo rating to take the difficulty of the matches into account. If a team plays five matches in a row against better teams, it is

harder to achieve just as many points as if a team plays against relativity worse teams. For the first five match rounds of each season the teams did not play five games yet, so for these match rounds the number of points of all previous matches in that season is used which lead to only missing values for the first match rounds. The missing values for the first match rounds are again set to zero.

The last variables which are scraped from the internet and added to the data set are the head-to-head ratio's (H2H). The website used for scraping is https://www.soccerbase.com/, which contains all historic match results in all different leagues between two teams. In this case all historic matches are considered instead of only the matches of the corresponding season to include some information about the rivalry between the teams. After scraping all match results of all teams which are in the data set, the head-to-head ratios are calculated. For a specific match, the head-to-head ratio is calculated by taking the proportional distribution of winners till the date of the match day wherefore there is no information in the data which is not known yet. These variables are in ratios to have the same scale as the ratios of home and away wins.

3.1.2. Descriptive Statistics

The data set is now complete, which means that the descriptive statistics of the data set can be investigated. The final data set contains 7,220 matches and 37 variables in total. There are five non-numeric variables, which are the date, the season, the name of the home team, the name of the away team, and the full-time result. The proportional distribution of home wins, draws, and away wins of the data set is shown in Table 2. From all 7,220 matches, roughly 46% of the matches are won by the home team, 29% by the away team, and 25% of the matches resulted in a draw. This implies that there is also a home field advantage in this date set. The bookmakers have an accuracy of 54.02% over the whole data set. This is a good indicator of how well the bookmakers are 'predicting' the matches. None of all the matches were predicted to result in a draw by the bookmakers. The intuition behind this could be that draws are less attractive for the bettors. The bettors are probably more interested in a winning (or losing) team. The bookmakers take this emotion into account in their odds.

Table 2: Proportional distribution of home wins, draws and away wins.

Home	Draw	Away
46.039%	24.861%	29.100%

The descriptive statistics of all numeric variables are shown in Table 3, which are 32 variables. The definitions of all variables are shown in Table A-1 in Appendix A: Variable Definitions. Table 3 shows that the mean of the odds for a home win (2.741) is lower than for a draw (3.934) or an away win (4.796). This means that also according to the bookmaker odds, in general the home playing team is more likely to win. The mean of the ratio of home wins of the home team (0.422) is also higher compared to the mean of the ratio of away wins of the away team (0.285). The lowest value for the Elo rating in this data set is 1474 and the maximum is 2084, where the mean is around 1718. Moreover, the statistics of the *Home Field Advantage* shows that this variable should not be a constant number because the minimum and maximum differ 82.3 points. The mean of this variable is about 63, which means that the proposed home field advantage of Hvattum & Arntzen (2010) would be a bit too high for this data set. A point to mention for the draw probability based on the Elo rating is that the maximum of this variable is about 14.676%, which means that this method still never predicts a match that results in a draw.

Table 3: Descriptive statistics of all numeric variables. The definitions of all variables are shown in Table A-1 in Appendix A: Variable Definitions.

Variable	N.	Mean	SD.	Min.	Max.
Attendance	7220	33473.506	16423.576	0.000	83222.000
Match Round	7220	19.500	10.967	1.000	38.000
Odds B365 Home Team Winning	7220	2.741	1.906	1.060	23.000
Odds B365 Draw	7220	3.934	1.136	2.500	17.000
Odds B365 Away Team Winning	7220	4.796	3.972	1.120	41.000
Rank Home Team	7220	10.499	5.959	0.000	20.000
Rank Away Team	7220	10.280	5.910	0.000	20.000
Attack Strength Home Team	7220	0.945	0.417	0.000	3.750
Attack Strength Away Team	7220	0.958	0.424	0.000	4.615
Defence Weakness Home Team	7220	0.960	0.383	0.000	4.615
Defence Weakness Away Team	7220	0.946	0.373	0.000	3.704
Average Points Home Team	7220	1.321	0.611	0.000	3.000
Average Points Away Team	7220	1.348	0.613	0.000	3.000
Average Losing Points Home Team	7220	1.593	0.630	0.000	3.000
Average Losing Points Away Team	7220	1.566	0.628	0.000	3.000
Average Goal Difference Home Team	7220	-0.019	0.835	-6.000	6.000
Average Goal Difference Away Team	7220	0.021	0.834	-5.000	6.000
Home Wins Ratio Home Team	7220	0.422	0.264	0.000	1.000
Away Wins Ratio Away Team	7220	0.285	0.245	0.000	1.000
Home Draws Ratio Home Team	7220	0.241	0.195	0.000	1.000
Away Draws Ratio Away Team	7220	0.241	0.198	0.000	1.000
Form Home Team	7220	2244.478	1256.129	0.000	5724.000
Form Away Team	7220	2339.860	1254.813	0.000	5931.000
Elo Rating Home Team	7220	1718.395	111.649	1496.000	2082.000
Elo Rating Away Team	7220	1718.524	111.642	1474.000	2084.000
Home Field Advantage	7220	63.085	13.118	17.500	99.800
Probability Home Team Winning (Elo)	7220	54.695	20.234	7.497	96.573
Probability Draw (Elo)	7220	5.820	5.412	0.000	14.676
Probability Away Team Winning (Elo)	7220	39.485	18.625	3.427	92.503
Win Ratio Home Team (H2H)	7220	0.374	0.115	0.000	1.000
Win Ratio Away Team (H2H)	7220	0.376	0.115	0.000	1.000
Draw Ratio (H2H)	7220	0.244	0.064	0.000	0.667

To give an overview of the performances of some teams of the Premier League over the past years, Figure 1 shows the development of the Elo rating for the 'big six' of the Premier League. The lines are smoothed using Generalized Additive Model (GAM) smoothing to have a better general overview of the development. The figure shows that Manchester City and Tottenham Hotspurs did not have a high Elo rating in the years 2002 till 2010 compared to the other clubs. From 2010 and onwards, the Elo ratings are more equal and in the last few years Liverpool and Manchester City had a higher Elo rating than their competitors which also resulted in the title for Manchester City in the seasons 2017/2018, 2018/2019, and 2020/2021 and for Liverpool in season 2019/2020. The rise of Manchester City in this graph can be explained by that fact that the club is purchased in 2008 by the Abu Dhabi United Group, so that the club received considerable financial investment to buy better players, which resulted in better performances.

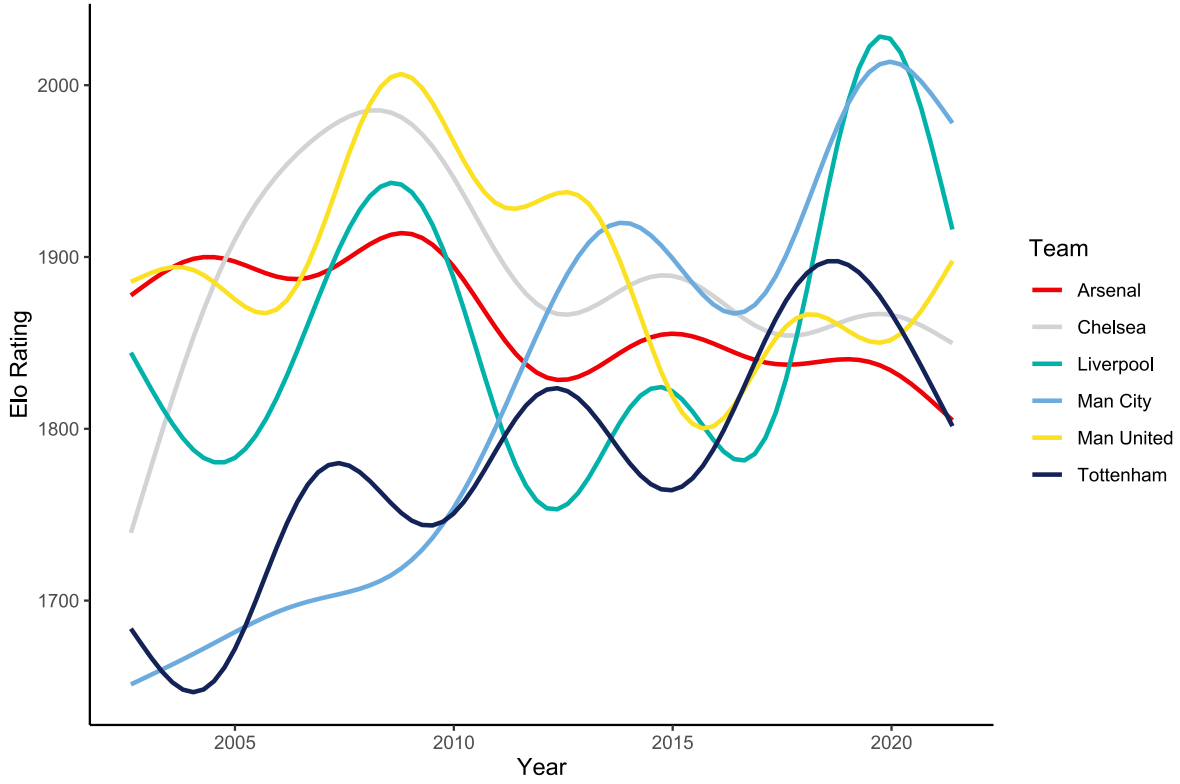

Figure 1: Development of the Elo rating for the 'big six' of the Premier League from season 2002/2003 up to and including season 2020/2021. Source: http://clubelo.com/.

Figure 2 shows the development of the *Home Field Advantage*, which should be added to the Elo rating of the home playing team for calculating the probabilities of winning, drawing, and losing based on the Elo rating. This figure is smoothed in the same way as Figure 1 for the same purpose. The figure shows that the *Home Field Advantage* did fluctuate over time but has a decreasing trend over the years. The Union of European Football Associations (UEFA) recently changed their rules about the home- and away goals in the European competition due the decrease in the home field advantage over the years (UEFA, 2021). Before the change of this rule, the away goals counted for two goals in the knock-out stages when the aggregated result was a draw. Figure 2 also shows that the *Home Field Advantage* has a large decrease since 2020. Due to COVID-19, most of the year of 2020 spectators were not allowed at the stadium which is the reason that the *Home Field Advantage* decreased. Nevertheless, the minimum of the *Home Field Advantage* is not zero which means there is always a home field advantage due to the field, travel time, or other aspects.

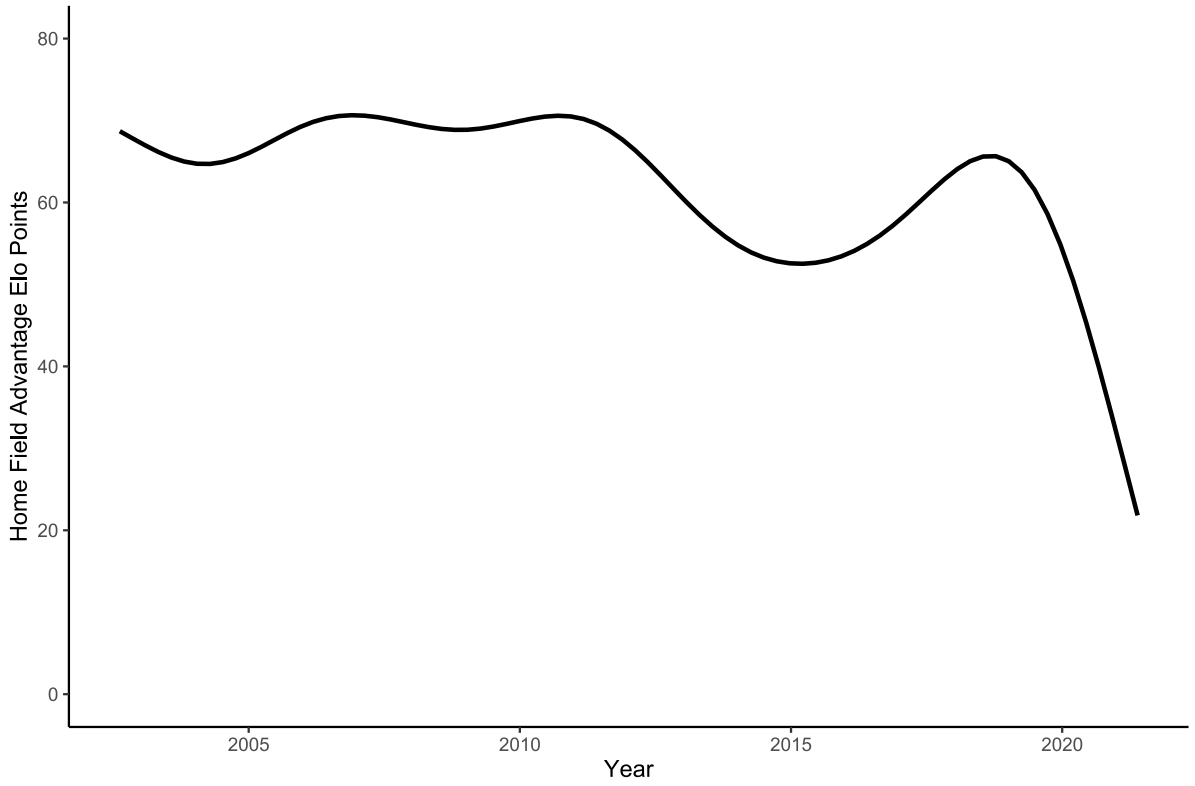

Figure 2: Development of the Home Field Advantage Elo Points from season 2002/2003 up to and including season 2021/2021. Source: http://clubelo.com/.

3.2. Methodology

All methods which are used for this research paper are part of machine learning. As already explained in previous section, machine learning is part of Artificial Intelligence (AI) and is the study of computer algorithms that improve automatically using data and through experience (Mitchell, 1997). This section will elaborate on multiple machine learning methods. Firstly, machine learning will be explained in general and afterwards all used machine learning methods for this research paper will be discussed in more detail.

Machine learning focuses on applications that learn from experience and improve their predictive accuracy and can make decisions with minimal human intervention. Artificial intelligence is the simulation of human intelligence in machines and the goals are learning, reasoning, and perception. Within machine learning there are three primary categories: supervised, unsupervised, and semi-supervised machine learning. With supervised machine learning the model trains itself based on data, which is labelled, whereas with unsupervised machine learning there is no need to supervise the model. Semi-supervised machine learning is a combination of both methods (Mohri, Rostamizadeh, & Talwalkar, 2018). For this research paper, supervised machine learning will be used because this research has to do with a

classification problem. The aim is to classify (label) a match to either the home team wins, the away team wins or a draw.

The first step of using machine learning is selecting and preparing the data. The data are already selected in the previous part of this section however the data are not yet prepared for usage. The data set should be divided into a training set and testing set. The training set is used to train the model and the testing set is used to measure the performance of the model. For this research, the first eighteen seasons are used as training set and the season 2020/2021 is used as testing set. After splitting the data, it is time to choose an algorithm to train the model. For this research, multiple algorithms are used to compare different results with each other. The used machine learning algorithms are the multinomial logistic regression, random forest, and Artificial Neural Network (ANN). All those machine learning methods will be described in more detail in the following paragraphs. When the algorithm is chosen, the algorithm can be trained with the training data. Training in machine learning is an iterative process, which means the repetition of a process in order to generate an outcome. The last step of the machine learning process is to use and improve the model. The trained model will be used on the test set which leads to a performance of the model on data where it is not trained on (Lantz, 2013).

3.2.1. Multinomial Logistic Regression

After the brief introduction of machine learning it is time to dive into some specific machine learning methods. The first machine learning method that will be described is the multinomial logistic regression, which is an extension of the binomial logistic regression. The dependent variable in the case of this research has three possible discrete outcomes, namely *Home*, *Draw* or *Away* and therefore a multinomial logistic regression is used. Nevertheless, the multinomial logistic regression has some assumptions which should be checked to make sure the final output is valid (Nerlove & Press, 1973). The first assumption is that the dependent variable must be nominal, which is the case with the full-time result. The second assumption is that all categorical independent variables should be converted to dummy variables. This assumption would mean that the eight non-numeric variables should be converted to dummies which leads to 2,517 variables. Due to the increase in computation time and no increase in accuracy, the categorical variables are left out. This means that the number of input variables is equal to 32. The third assumption is that there should be no outliers or high influential

points in the data. Based on the Chi-squared test, Cochran test, Dixon's Q-test, and Grubbs' test, all outliers are detected and replaced by the mean of the predictor (Komsta, 2011). In total, there were 2.333 outliers in the training data set, which is about 1.07% of all values. The fourth assumption is that there should be a linear relationship between the dependent variable and continuous independent variables. This cannot be measured directly and therefore the logit transformation is taken from the dependent variable ($logit\ p\ = \log\left(p/(1-p)\right)$). The final assumption is that there should be no multicollinearity. This assumption is tested by calculating the variance inflation factor (VIF) for each variable. All variables which have a VIF higher than ten are removed as predictor for the multinomial logistic regression, meaning that there are 17 predictors left for the multinomial logistic regression (Fox & Weisberg, 2019). This threshold of ten is known as a rule of thumb when using the VIF (Craney & Surles, 2002).

3.2.2. Random Forest

The second machine learning method which will be described is the random forest. The random forest algorithm averages multiple deep decision trees which are trained on different parts of the same training data. The goal of this algorithm is to overcome the over-fitting problem of individual decision tree. In other words, the random forest is an ensemble learning method and will be used as classifier in this research. For a classification problem, the algorithm uses multiple decision trees to give the mode of the classes as output.

The random forest algorithm starts with the random record selection. Each decision tree is trained on a proportion of the total training data (63.2%). The remaining data are called the out-of-bag (OOB) (Rosenberg, 2017). The proportion of 63.2% is drawn at random with replacement from the original data. The second step of the algorithm is to select random variables. Some independent variables are selected random, and the best split of those selected variables are used to split the node. The other data, called out-of-bag (OOB), is used to calculate the misclassification rate (OOB error rate). The previous steps will be repeated $ntree$ times, where $ntree$ is the number of trees. Each tree then gives a classification based on the OOB. Based on a majority vote, the algorithm chooses the prediction. The probability of class will be calculated by dividing the number of votes for that class by the total number of votes. 'Random' in random forest refers to the random observations to grow each tree and to random selecting variables for splitting at each node. (Friedman, Hastie, & Tibshirani, 2001). Because

the OOB error rate for the model with the non-numeric variables was higher than the OOB error rate for the model without the non-numeric variables, the non-numeric variables are also excluded for the random forest.

The two parameters which can be tuned for a random forest are the number of trees ($ntree$) and the number of random variables used in each tree ($mtry$). By default, $mtry$ is square root of the total number of all predictors. Before tuning $mtry$, the optimal value for $ntree$ needs to be determined. The $ntree$ is set to 750 based on trial and error, with as goal to find the value where the OOB error rate reaches a minimum. The $mtry$ is set to the default, which is six ($\sqrt{32} \approx 6$), because the OOB error rate did not decrease when trying other values.

3.2.3. Artificial Neural Network

There are many types of neural networks, however in this research the Artificial Neural Network (ANN) will be used instead of the Recurrent Neural Network (RNN). ANN is also known as a feedforward neural network and a RNN is also known as a feedback neural network. The reasoning behind the choice for the ANN is that the data set already contains much information about the previous match rounds or the form of the team. Therefore, it is not necessary to use an RNN.

A neural network is a complex adaptive system, which means that the network has the ability to change its internal structure by adjusting the input weights. Thereby, the neural network is an information processing model and can learn from examples. The structure of a neural network can be seen as the human brain; it has a large number of highly interconnected processing elements, which are known as the neurons. Throughout the neurons, it follows the non-linear path and processes information. Originally, the neural network was designed for pattern recognition, however in this research paper it will be used as classifier (Anthony & Bartlett, 2009).

The neural network has an input layer, one or more hidden layers, and an output layer. Each layer has one or multiple neurons, where for the input and output layer the number of neurons is equal to the number of input and output variables, respectively. The input variables have weights and biases, which are adjustable parameters. These parameters can be adjusted using some learning rules. However, the output of a neuron can take every number and has no

boundary. Therefore, an activation function is used as mapping mechanism between the input and output of the neuron. There are multiple types of activation functions, however in this research the Rectifier Linear Unit (ReLU) activation function is used which is the most used activation function in neural networks. However, this is not the reason for choosing this activation function. The choice for this activation function is based on the prediction accuracy. This activation function had the highest prediction accuracy. The equation for the ReLU activation function is as follows: $f(x) = \max(0, x)$. If the output node is negative, it will be set to zero.

In comparison to the logistic regression and the random forest, there are many more parameters to tune with a neural network. The first parameters which are tuned are the number of hidden layers and the number of hidden nodes. After trying many combinations of hidden layers and hidden nodes, the network with just one hidden layer and one hidden node is performing the best. The computation time of this model is also much less than for networks with many hidden layers and hidden nodes. As optimization algorithm, gradient descent is used and especially the RPROP+. RPROP is short for resilient backpropagation and is created by Riedmiller & Braun (1992). The RPROP+ algorithm refer to the resilient backpropagation with weight backtracking. The non-numeric variables are not included in the neural network due to the increase in input nodes because of all added dummies and consequently the computation time, meaning that the input nodes are all numeric variables.

3.2.4. Performance Measures

To evaluate the performance of all three machine learning models and to give an answer to the first sub-question, two performance measures will be used. The first performance measure is the classification accuracy, which is also used in all previous research papers. All machine learning models give predictions for each match which will be compared to the actual results of the matches. The number of correct predictions divided by the total number of matches is the accuracy. The accuracy of the bookmaker Bet365 is for the test set is 51.84%, which is lower than the accuracy of the total data set (54.02%), meaning that the bookmakers did relatively worse in predicting compared to the previous seasons. This accuracy will be used as benchmark for evaluating the performance of the machine learning models.

Nevertheless, this research paper is not only about getting the highest accuracy but more about getting the highest return (or making the most profit). Previous research papers focus only on the accuracy and barely on other measures. Therefore, this research paper also uses the return on investment (ROI) as performance measure. The return on investment is a percentage of how much profit is made divided by the original investment. For calculating the profit of a bet, the maximum market closing odds are used. The maximum market closing odds are the highest odds in the market just before the start of a match. This means that the odds are compared among different bookmakers and the most interesting odds are used. This data are also gathered from www.football-data.co.uk/data.php. For this research paper, the return on investment will be slightly more important than the accuracy of the algorithms because this research paper is about researching the best betting strategy. The return on investment is also used to compare the performances of the different betting strategies and money management techniques among each other.

To have some sort of distribution of the return on investment, the development of the bankroll will be plotted in the results part. The development of the return on investment and the bankroll can differ because the development of the return on investment is relative, where the development of the bankroll is absolute. Furthermore, the percentage of number of bets won is also relevant as performance measure to see how many bets are won by the different money management techniques, however the return on investment will be the most important measure.

3.2.5. Variable Importance

To answer the second sub-question, it is necessary to have a variable importance overview. Because it is not yet known which machine learning model is performing the best, the methodology for calculating the variable importance for all models will be discussed. The variable importance is relative among the predictors and cannot compared between the different models. Consequently, the absolute values of the variable importance cannot be interpreted.

For the logistic regression, the absolute value of the t-statistic for each model parameter is used to calculate the variable importance. The parameter with highest absolute t-value, is the most important predictor in the model (Kuhn, 2020).

The calculation for the random forest is a bit more complicated. There is a prediction accuracy of the out-of-bag sample, and a prediction accuracy after each variable is permuted. The difference between those accuracies is normalized by the standard error and averaged over all trees. The higher this score, the more important the predictor is in the model.

The connection weights in an artificial neural network are similar to the coefficients in a logistic regression, however there many more connecting weights in a neural network than coefficients in a logistic regression. The interpretation of those connection weights is thus much harder than for the coefficients in a logistic regression. Therefore, Olden, Joy, & Death (2004) proposed an algorithm which can calculated the relative variable importance. This method is called the Olden method and calculates the variable importance as the raw input-hidden and hidden-output connection weights between each input and output node. After that, the products across all hidden nodes are summed. The higher this score, the more important the predictor is in the neural network.

3.2.6. Betting Strategies

Two different betting strategies and three money management techniques will be compared to each other in this research paper. The first betting strategy will be to bet on each match based on the labels calculated by the algorithms. The second betting strategy will use the bookmakers' odds to choose on which matches to bet (Caan Berry Pro Trader, 2020). The probabilities calculated by the algorithms can be converted to odds. This conversion can be done by dividing one by the probabilities ($1 \ / \ prob$). This betting strategy will only bet on matches which are mispriced by the bookmakers. If the odds are higher at the bookmaker compared to the calculated odds, it means that the match is mispriced. This means that the potential pay-out is higher than it should be. With the second betting strategy, you place a bet on the team which has the highest difference between the bookmaker odds and the created odds, where the bookmaker odd should be higher. These two betting strategies will be compared to each other to see if there is a difference between betting on each match of picking some matches to bet on.

These two betting strategies are based on single betting, which means that the bettor places a bet on only one match at a time. The bettor could also bet on more than one match at a time, which is known as combination betting. The pay-out for combination betting is

higher because the risk is also higher and is calculated by multiplying the odds of all matches with each other. For example, if you place a bet on a match with an odd of 1.5 and on a match with 1.7, the odd for the total bet will be $1.5 \times 1.7 = 2.55$. This means that combination betting could lead to higher pay-outs, however this strategy is not used in this research. The reasoning behind this is based on Rue & Salvesen (2000), because they conclude in their research paper: "It seems to be both easier and more realiable to bet on single matches compared with combination bets."

3.2.7. Money Management Techniques

Within those betting strategies there are also different money management techniques, which are already introduced in the literature review. The money management strategies which will be compared to each other in this research are fixed betting, proportional betting, fixed expected return, the Kelly criterion, the Fibonacci sequence (with three forms), and variance-adjusted betting. All those eight money management techniques are already explained in the last part of the Literature Review; meaning that these techniques will not be explained again but the implementation of the different techniques will be discussed in this part.

The first money management technique is the fixed betting strategy, where you bet the same amount on each match. The fixed stake for each bet is ten units, where the total stake – original investment – is thus ten times the number of betted matches. Proportional betting is the second money management technique and the proportion which is used is 1% of the current bankroll, where the starting bankroll is one hundred units. This means that on the first match one unit is betted. The third money management technique is the fixed expected return strategy. This strategy tends to have the same expected return for all matches. The expected return to achieve is set to five units. With the Kelly criterion, you bet a proportion of your current bankroll on each match, however this proportion is calculated based on the probability of winning (see equation (1)). The starting bankroll for this strategy is the same as for the proportional betting technique and is equal to one hundred units. The Fibonacci sequence is used as fifth money management technique. If you win a bet, you move down two units in the sequence and place that number of the sequence on a bet. If you lose, you move up one unit in the sequence. This technique has a limitation that it leads to really large stakes for bets which are probably not realistic within the current betting market. Therefore, this

technique will use three different forms: an unlimited form, a form with a limit of 1.000 units per bet, and a form with a limit of 100 units per bet. The last money management technique is the variance-adjusted technique, which tries to minimize the difference between the expected profit and the variance of that profit. The amount of stake for each bet is calculated by equation (2). All those different money management techniques and the two betting strategies will be compared to each other using the return on investment as measure.

In summary, there are two betting strategies and eight money management techniques. The different betting strategies are used to see if there is a difference between betting on each match of picking some matches to bet on. The different money management techniques are used to see which way of allocating your money over all the matches is the most profitable. In the next section the results of both the betting strategies and money management techniques will be discussed.

4. Results

This section of the research paper will present the results of the multiple algorithms. As explained in the previous section, the accuracy and the return on investment will be used as measure for the performance of the different algorithms but also for the different betting strategies and money management techniques. The test set will be used for evaluating the performance of the algorithms.

4.1. Prediction Accuracy

To evaluate the accuracies of the machine learning methods, it is good to have some sort of benchmark. The accuracy of the bookmakers will be used as benchmark, which is 54.14% for the training set and 51.84% for the test set. The benchmark for the test set is lower than the accuracy of the total data set (54.02%), which means that the bookmakers did relatively worse in predicting compared to the previous seasons. A possible reason could be that season 2020/2021 differ compared to the other seasons due the different circumstances because of COVID-19 measures. The proportion distribution of the number of home wins, draws and away wins is also a bit different. Actually, most of the matches were won by the away playing team in this season (40.26%). Around 21.84% of the matches resulted in a draw and the other 37.89% of the matches were won by the home playing team. This means that there was less home field advantage in this season which is confirmed by Figure 2. The return on investment based on the bookmaker odds is for all different money management techniques negative, except for the Kelly criterion. The benchmark for return on investment is to be at least as profitable as the Kelly criterion based on the bookmaker odds.

The first measure that will be compared with each other is the prediction accuracy. The prediction accuracy is the percentage of how many bets you will win if you place a bet on each match. The multinomial logistic regression has a prediction accuracy of 54.18% and 54.74% for the train and test set, respectively. These two percentages differ not that much from each other, which means that is no underfitting or overfitting. The multinomial logistic regression outperforms the bookmakers' accuracy for the test set, which offers opportunities for being profitable. The logistic regression predicts that most of the matches are won by the away playing team (232) and just one match were predicted as a draw. The rest of the matches were predicted to be won by the home playing team (147).

The random forest has a prediction accuracy of 52.11% for the test set. The accuracy of the training set is based on the OOB-error, which is 47.28%. The accuracy of the training set is therefore $100\% - 47.28\% = 52.72\%$, meaning that also with this model there is no underfitting or overfitting. The random forest algorithm is outperforming the bookmakers' accuracy for the test set, however not for the training set. The random forest predicts that more matches are won by the home team than the away team, which is not the case in the season of 2020/2021. The algorithm predicts admittedly more draws than the logistic regression, namely seventeen draws. In conclusion, the random forest is performing worse in comparison with the logistic regression.

The artificial neural network has a prediction accuracy of 54.44% and 54.47% for the train and test set, respectively. This algorithm is outperforming the bookmakers' accuracy the training set and test set. The prediction accuracies are approximately equal to the prediction accuracies for the multinomial logistic regression; however, the artificial neural network is performing a bit better in the training set, but the logistic regression is performing a bit better in the test set. The neural network does not predict any draws but does predict more away wins (213) than home wins (167). Based on this performance measure, the main research question and the first sub-question cannot be answered, because the research question is about to what extent it is possible to be profitable and not just about how accurate the model can predict. To see if there is an opportunity to be profitable using the predictions of these machine learning models and to have a clearer overview of which machine learning models is performing better, the next performance measure will be discussed which is the return on investment.

4.2. Return on Investment

Firstly, the results of the first betting strategy (bet on each match) will be discussed. This betting strategy is to bet on each match based on the labelled class of the different machine learning models. The results of this betting strategy and the different money management strategies are shown in Table 4. The bold prediction accuracies are the machine learning model which is performing the best for the specific money management technique. The maximum market closing odds are used as odds to calculate the return on investment. The table shows that the prediction accuracy of the bookmakers for all money management techniques are negative, except for the Kelly criterion. Using the bookmaker odds, it is only

possible to be profitable when using the Kelly criterion, however this profit is really low. Nevertheless, it still the only profitable technique and for that reason it is probably popular among the bettors.

Table 4: Return on investment for betting on each match.

Money Management Technique	Logistic Regression	Random Forest	Neural Network	Bookmakers
Fixed Betting	**11,12%**	3,78%	10,22%	-4,57%
Proportional Betting	**10,78%**	3,18%	10,16%	-4,97%
Kelly Criterion	**4,82%**	1,91%	3,44%	0,20%
Fixed Expected Return	9,66%	3,95%	**9,75%**	-4,56%
Fibonacci Sequence (Unlimited)	**14,90%**	6,20%	12,72%	-9,34%
Fibonacci Sequence (Limit 1000)	**14,90%**	6,20%	12,72%	-9,34%
Fibonacci Sequence (Limit 100)	**14,90%**	6,20%	12,72%	-13,43%
Variance-Adjusted	0,97%	-0,66%	**1,38%**	-4,51%

Thereby, the table shows that for all money management techniques, excluding the fixed expected return and variance-adjusted technique, the logistic regression has the highest return on investment. The best performing money management technique is the Fibonacci Sequence for all machine learning models. There is no difference between the three forms of the Fibonacci sequence, meaning that there was no stake which exceeded the limit of 100 units. Based on this table, it can be concluded that for the first betting strategy the logistic regression is the best performing machine learning model, and the Fibonacci Sequence is the best performing money management technique.

The results of the second betting strategy (bet on misprised teams) are presented in Table 5. Compared to Table 4, all the returns on investment are higher except for the Kelly criterion. For all other money management techniques, this means that the second betting strategy is performing better than the first betting strategy. Only for the Kelly criterion, one could better use the real odds of the bookmakers instead of the 'true' odds.

Table 5: Return on investment for betting on the teams which are mispriced by the bookmakers.

Money Management Technique	Logistic Regression	Random Forest	Neural Network	Bookmakers
Fixed Betting	**28,61%**	21,28%	25,72%	-4,97%
Proportional Betting	**28,89%**	18,77%	21,40%	-4,97%
Kelly Criterion	**11,80%**	-7,60%	2,53%	0,20%
Fixed Expected Return	**24,63%**	20,02%	24,04%	-4,56%
Fibonacci Sequence (Unlimited)	60,97%	**114,12%**	89,32%	-9,34%
Fibonacci Sequence (Limit 1000)	**60,97%**	22,71%	21,02%	-9,34%
Fibonacci Sequence (Limit 100)	**60,27%**	21,12%	16,84%	-13,43%
Variance-Adjusted	7,58%	11,37%	**24,59%**	-4,51%

In line with Table 4, the logistic regression has the highest returns on investment for the first four money management techniques. The random forest is performing the best for the money management technique the Fibonacci sequence and the neural network for the variance-adjusted technique. The returns on investment for the Fibonacci sequence technique with an

unlimited stake are relatively high compared to the other techniques. The total stake for this technique which is needed for those returns are also very high. The total stake for the random forest for example is more than a trillion units. Besides the fact that the bookmakers do not allow bettors to place bets of extremely high amounts, it is also unlikely to have that much amount of money. The table shows that a limit of the stake leads to lower returns on investment, but also a lower total stake. There is no difference in return on investment between the Fibonacci sequence for the logistic regression with the unlimited stake and a limit of 1.000 units. This means that the logistic regression never has a single stake of more than 1.000 units. The total stake with both forms is 4.025 units. The total stake for the form with a limit of 100 units is about 100 units less than for the other forms but the return on investment is also a bit lower. Because a limit of 1.000 units per bet is still a large amount of money to bet, one would consider the form with a limit of 100 units to use in practise. A disadvantage of this form is that the concept of the Fibonacci sequence is less strong because the limit is much lower. If the bettor loses a couple of bets in a row and the original stake is multiple times above 100 units, it is more difficult to earn the lost money back.

The second-best performing money management technique for the logistic regression is the proportional betting technique, which is almost equal to the fixed betting technique. The results in Table 4 and Table 5 are specific for just one season. Therefore, the best performing money management technique with the logistic regression predictions will be applied to four other seasons to see which technique gives higher returns in general. The best performing techniques are the fixed betting technique, proportional betting technique, and Fibonacci sequence with a limit of 100 units. Table 6 shows the returns on investment for the seasons 2016/2017 up to and including season 2020/2021 for those three techniques. The model is trained again on all previous seasons. The returns are calculated with the bookmaker odds of Bet365 because the maximum closing odds data are not available for all seasons. Table 6 shows that the returns on investments differ per season; some seasons have negative returns, and some seasons have positive returns. The table shows that season 2020/2021 is the most profitable season. In most seasons the Fibonacci sequence is outperforming the other two money management techniques and the average return on investment over the five seasons is also higher. Based on Table 6, it can be concluded that the Fibonacci sequence is the best

performing money management technique is with the logistic regression as machine learning model.

Table 6: Returns on investments for different seasons for the best performing money management techniques.

Season	Fixed Betting	Proportional Betting	Fibonacci Sequence (Limit 100)
2016/2017	-9,50%	-10,21%	-1,99%
2017/2018	1,04%	0,12%	-4,07%
2018/2019	17,27%	15,30%	28,04%
2019/2020	-2,30%	-3,58%	3,61%
2020/2021	22,79%	22,57%	50,73%
Average	5,86%	4,84%	15,27%

The development of the return on investment for the Fibonacci sequence with a limit of 100 units for the second betting strategy (bet on misprised teams) is plotted in Figure 3. The figures for the development of the returns on investment and the bankrolls for the second betting strategy and all money management techniques are shown in Appendix B: Graphs. Figure 3 shows that the development of the return on investment for the logistic regression is outperforming all other models and is quite constant. Despite the fact that the Fibonacci sequence is in theory the best performing technique, this technique is riskier than the other techniques because the amount of stake increases when losing bets. Even with a limit of 100 units, the total amount of stake can increase to really large amounts. For example, the total stake for season 2016/2017 is almost 11.000 units and for season 2019/2020 around 18.500 units. Based on these results, it can be concluded that the Fibonacci sequence is not the most rational and realistic technique. The fixed betting and proportional betting techniques are more rational, where the fixed betting technique is performing a bit better than the proportional betting technique based on Table 6.

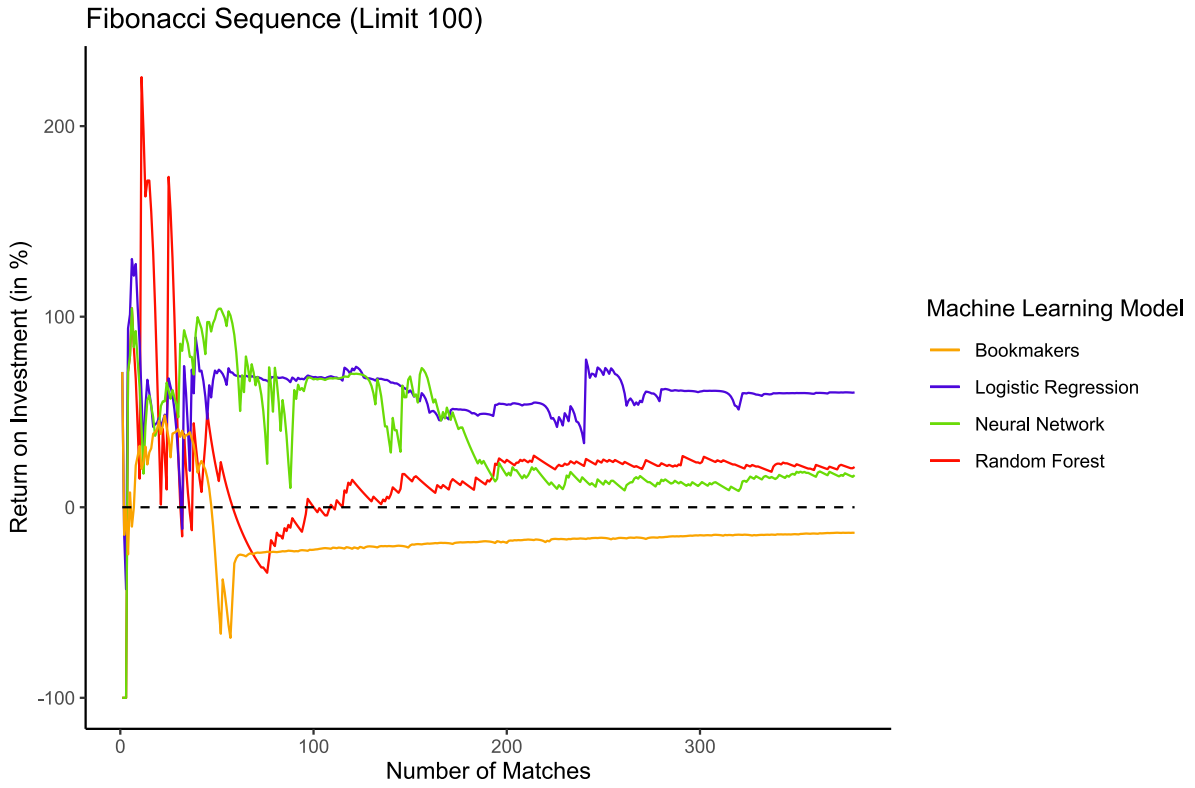

Figure 3: Development of the return on investment for the second betting strategy (bet on misprised teams) with the Fibonacci sequence with a limit of 100 units.

Based on Table 4 and Table 5, it can be concluded that the second betting strategy (bet on misprised teams) is outperforming the first betting (bet on each match) strategy for all money management techniques except for the Kelly criterion. This means that it is more profitable to bet only on the matches which are mispriced by the bookmakers. When looking at the best performing machine learning model, it can be concluded that the (simple) logistic regression is performing the best, based on the returns on investment in Table 5. Based on this conclusion and the prediction accuracy, the first sub-question can be answered. The answer to this first sub-question is that the logistic regression performs the best in predicting football match results in terms of returns on investment. Based on the prediction accuracy, the logistic regression and the neural network differ not that much.

When using the logistic regression and excluding the Fibonacci sequence due to the risk, the best money management technique is the fixed betting technique, which is performing a bit better than the proportional betting technique (see Table 6). The simplest money management techniques are thus more profitable than the more advanced money management techniques. The answer to the third sub-question is that the fixed betting technique is the most profitable technique when using the logistic regression as machine learning model.

The second betting strategy (bet on misprised teams) with the fixed betting technique based on the predictions of the logistic regression is thus the best performing model. To compare the development of the return on investment of the logistic regression with this strategy with the other machine learning models, Figure 4 is plotted. This figure shows that all machine learning models are outperforming the bookmakers and that for the logistic regression and the neural network the return on investment is only for the first few matches negative. The returns on investment for the first few match rounds are probably negative because not all information is yet available and a couple of lost bets in the early stages lead to negative returns. This figure also shows that the neural network is performing better in the first half of the season. From around the 150th match the return on investment is decreasing. The development of the logistic regression is more constant; meaning that it is thus less profitable in the first half of the season.

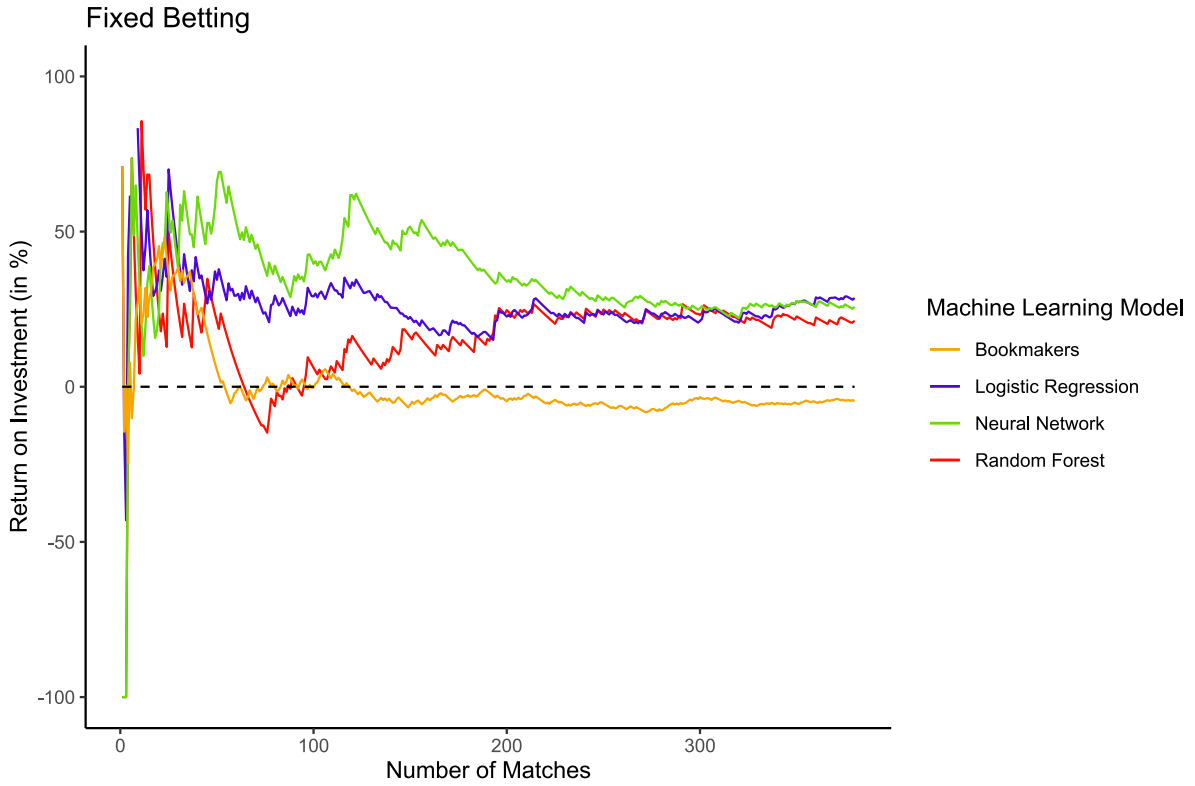

Figure 4: Development of the return on investment for the second betting (bet on misprised teams) strategy with the fixed betting technique.

The return on investment can be increased by combining the neural network and logistic regression for season 2020/2021. The best cut-off to switch from the neural network to the logistic regression is after 169 matches, which lead to a return on investment of 34.02%. This means that this hybrid form of the machine learning models has a higher return on

investment of more than 5 percent points. Nevertheless, this cut-off is specific for this season and the optimal cut-off for other seasons are different. This means that a combination of more machine learning models cannot be used in practise and will thus not be the best performing model.

In conclusion, the best performing model is the second betting strategy (bet on mispriced teams) with the fixed betting technique and the logistic regression as machine learning model. One would be tempted to use the winning strategy of this research paper. Based on the results, this will most likely result in profits of around six percent per annum (see Table 6). For season 2020/2021, this model and strategy has led to a return on investment of even 28.61%. The development of the bankroll of the fixed betting technique is plotted in Figure 5. This figure shows that the logistic regression leads to the highest bankroll in the end and the development is quite linear. With a total stake of 3680 units, a profit of 1052.9 units is made. In total, 149 bets of 368 bets are won, which is a win percentage of 39.21%. The win percentage is lower than prediction accuracy, however the profit which is made with the bets on teams which were mispriced are higher than the amount of money which is lost by the losing bets.

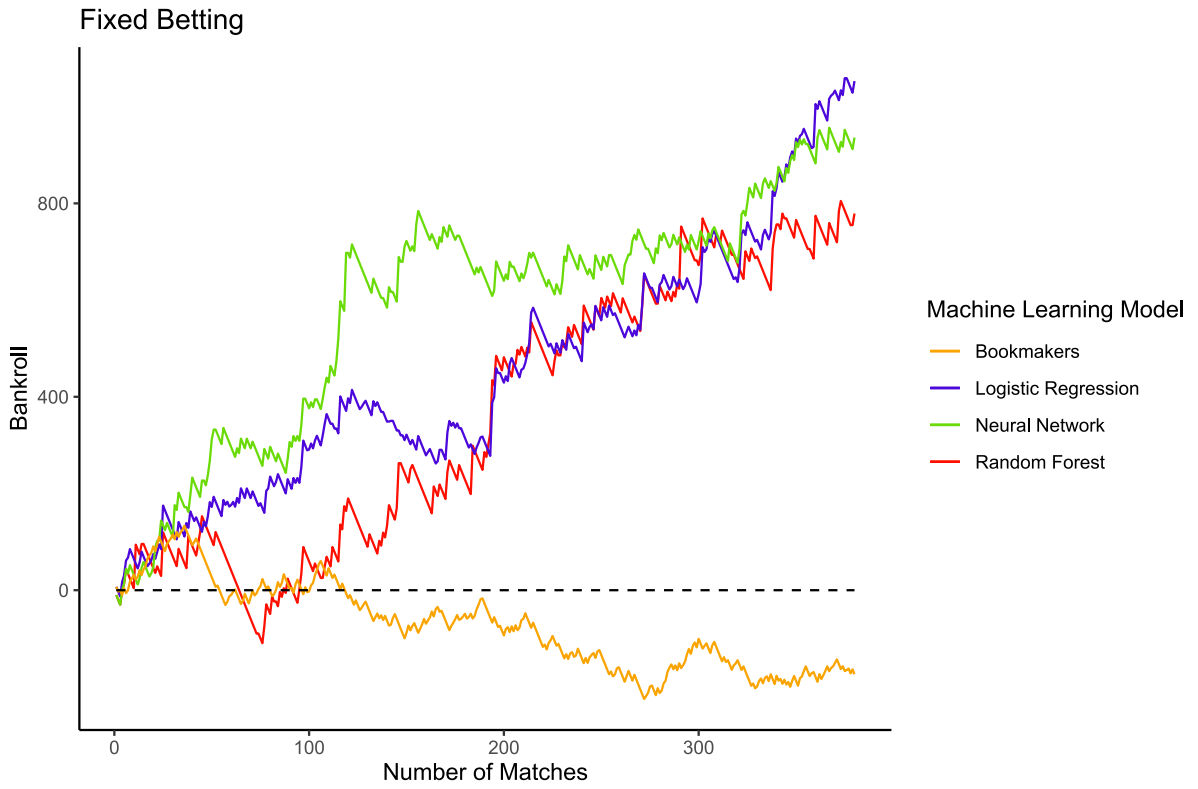

Figure 5: Development of the bankroll for the second betting (bet on misprised teams) strategy with the fixed betting technique.

4.3. Variable Importance

The best performing machine learning model is the multinomial logistic regression. To see which predictor is the most important is this model, the variable importance of this model needs to be calculated.

For the logistic regression there are less predictors than for the other models because of the deletion of some predictors based on the VIF. The most important predictors for the logistic regression are based on the absolute value of the t-statistic. The predictors with the highest absolute t-statistic are the *WinRatioH2HHomeTeam*, *DrawRatio*, and *WinRatioH2HAwayTeam*, which are thus the most important predictors. This means that the logistic regression learns the most from the head-to-head results. Besides these three variables, the betting odds of the bookmaker; the percentage of away wins of the away team; and the percentage of home wins of the home team are important, however these predictors are less important than the head-to-head results.

The second sub-question is as follows: *"Which predictor(s) has/have the most effect on predicting football match results in the introduced model of this research?"*. The answer to this question is that head-to-head results are the most important predictors in the introduced model. It can be concluded that the logistic regression learns the most from the head-to-head results.

5. Conclusion & Discussion

5.1. Conclusion

The first part of this section will conclude this research paper. The main research question is as follows: *"To what extent is it possible to be profitable in the football betting market using machine learning?"*. Before giving an answer to this question, all the sub-questions will be answered.

The first sub-question is about which machine learning will perform the best for predicting football match results. Three different machine learning models are compared to each other in this research paper, namely the multinomial logistic regression, random forest, and artificial neural network. The prediction accuracy of the test set for the logistic regression is the highest, however this prediction accuracy does not differ much from the accuracy of the neural network. When looking at the return on investment over all different money management techniques, it can be concluded that the logistic regression is performing better than the other two machine learning models. Therefore, the answer to the first sub-question is that the logistic regression is the best performing machine learning model for predicting football match results. This is partially in line with the results of previous research papers which concluded that the logistic regression and neural networks lead to the best results.

The second sub-question is about which predictor has or have the most effect on predicting football match results in the introduced model. This introduced model in this research paper is the logistic regression. The effect of the predictor is calculated by the variable importance. The variable importance for the logistic regression is based on the absolute value of the t-statistic and the most important predictors are the head-to-head results followed by the betting odds of the bookmaker and the percentage of home/away wins for the home/away team. Based on this variable importance, it can be concluded that the head-to-head results are the most important predictors in the introduced model.

The last sub-question of this research paper is which money management technique in football betting generate the most profit with the introduced model. The first thing to mention is that the second betting strategy (bet on misprised teams) is overall performing better than the first betting strategy (bet on each match). This means, that one can better place bets on teams which are mispriced by the bookmakers. Nevertheless, this does not answer the last sub-

question because that is about the money management techniques. Based on Table 5 and after eliminating the Fibonacci sequence due to risk, it can be concluded that the 'simple' money management techniques are more profitable than the more advanced techniques. The best performing money management technique is the fixed betting strategy when using the best performing machine learning model. In conclusion, the answer to the last sub-question is that the fixed betting technique is the most profitable.

The answer to the main research question is that it is possible to be profitable in the football betting market using machine learning. The betting strategy where you place bets on teams which are mispriced by the bookmakers are more profitable than the strategy to bet on every team which is predicted by the machine learning models. The simpler money management techniques are thereby more profitable than the more advanced money management techniques. The introduced model in this research paper will most likely result in profits of around six percent per annum. Seasons 2020/2021 has even a return on investment of 28.61% with the introduced model.

5.2. Discussion

Beside the fact that this research paper has led to an algorithm which is profitable, there are some limitations and ideas for further research. This part of the section will elaborate on this topic and will discuss the research paper in total.

The prediction accuracy of this research for the logistic regression is about 55% which is higher than the bookmakers' accuracy. However, as discussed in the Literature Review there are other research papers which managed to get higher prediction accuracies. Comparing the prediction accuracy of this research with the other accuracies, the prediction accuracy of this research is relatively low. Despite this accuracy, the proposed betting strategy and money management techniques have led to profitable results. Nevertheless, these results could perhaps be improved when having a higher prediction accuracy. More relevant predictors could be added or irrelevant predictors could be removed to try to improve the prediction accuracy.

Another limitation of this research paper is that it is specific for the English football competition and is not tested on other leagues. If the data is available, this model can easily tested on other leagues. An idea for further research could be to test the model on other leagues and on other seasons.

The returns on investment per season differ quite a bit per season. A limitation of this fact is that the model does not give a constant return per season. One season the return can be very high and the other season the return can be negative. The ideal scenario is to achieve a constant return each season. Further research could elaborate on the research of achieving a more constant return over the years.

References

Alfredo, Y. F., & Isa, S. M. (2019). Football match prediction with tree based model classification. *International Journal of Intelligent Systems and Applications, 11*, 20–28.

Aloufi, S., & El Saddik, A. (2018). Sentiment identification in football-specific tweets. *IEEE Access, 6*, 78609–78621.

Anthony, M., & Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations.* cambridge university press.

Brandessence Market Research & Consulting Pvt ltd. (2020). *Online Gambling Market.* London: Brandessence Market Research & Consulting Pvt ltd.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied computing and informatics, 15*, 27–33.

Buursma, D. (2011). Predicting sports events from past results Towards effective betting on football matches. *Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland*, *21.*

Caan Berry Pro Trader. (2020, 3). Betting Strategy That Works — Make an Income Betting on Sports. *Betting Strategy That Works — Make an Income Betting on Sports*. Retrieved from https://www.youtube.com/watch?v=iL4rmbwFwEY&list=WL&index=2

Constantinou, A., & Fenton, N. (2013). Profiting from arbitrage and odds biases of the European football gambling market. *Journal of Gambling Business and Economics*.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering, 14*, 391–403.

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*, 265–280.

European Gaming & Betting Association. (2020). *European Online Gambling – Key Figures 2020 Edition.* Brussels: European Gaming & Betting Association.

FIFA. (2018). *2018 FIFA World Cup Russia.* Paris: FIFA.

Forrest, D., & Simmons, R. (2001). *Globalisation and efficiency in the fixed-odds soccer betting market. University of Salford.* Tech. rep., Salford: Mimeo.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third ed.). Thousand Oaks CA: Sage. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Fritsch, S., Guenther, F., & Wright, M. N. (2019). *neuralnet: Training of Neural Networks.* Retrieved from https://CRAN.R-project.org/package=neuralnet

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert, 6*, 46–51.

Godin, F., Zuallaert, J., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2014). Beating the bookmakers: leveraging statistics and Twitter microposts for predicting soccer results. *Workshop on Large-Scale Sports Analytics (KDD 2014), New York, USA.*

Gray, P. K., & Gray, S. F. (1997). Testing market efficiency: Evidence from the NFL sports betting market. *The Journal of Finance, 52*, 1725–1737.

Hucaljuk, J., & Rakipovic, A. (2011). Predicting football scores using machine learning techniques. *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*, 460–470.

Igiri, C. P. (2015). Support Vector Machine–Based Prediction System for a Football Match Result. *IOSR Journal of Computer Engineering (IOSR-JCE), 17*, 21–26.

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

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems, 19*, 544–553.

Kampakis, S., & Adamides, A. (2014). Using Twitter to predict football outcomes. *arXiv preprint arXiv:1411.1243*.

Karlis, D., & Ntzoufras, I. (1998). Statistical modelling for soccer games: the greek league. 541–548.

Kelly Jr, J. L. (2011). A new interpretation of information rate. In *The Kelly capital growth investment criterion: theory and practice* (pp. 25–34). World Scientific.

Komsta, L. (2011). *outliers: Tests for outliers.* Retrieved from https://CRAN.R-project.org/package=outliers

Kuhn, M. (2020). *caret: Classification and Regression Training.* Retrieved from https://CRAN.R-project.org/package=caret

Langseth, H. (2013). Beating the bookie: A look at statistical models for prediction of football matches. *SCAI*, (pp. 165–174).

Lantz, B. (2013). *Machine learning with R.* Packt publishing ltd.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*, 18-22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports, 5*.

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

Markowitz, H. M. (1952). Portfolio Selection. *The Journal of Finance, 7*, 77. doi:10.2307/2975974

Mitchell, T. M. (1997). *Machine Learning* (1 ed.). USA: McGraw-Hill, Inc.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning.* MIT press.

Nerlove, M., & Press, S. J. (1973). *Univariate and multivariate log-linear and logistic models* (Vol. 1306). Rand Santa Monica.

Newall, P. W. (2015). How bookies make your money. *Judgement and Decision Making, 10*, 225–231.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological modelling, 178*, 389–397.

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*, 812.

Pettersson, D., & Nyquist, R. (2017). Football match prediction using deep learning. *Psychol. Sport Exerc., 15*, 538–547.

Prasetio, D., & Harlili, D. (2016). Predicting football match results with logistic regression. *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, (pp. 1-5). doi:10.1109/ICAICTA.2016.7803111

Rahman, M. A. (2020). A deep learning framework for football match prediction. *SN Applied Sciences, 2*, 1–12.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *IEEE international conference on neural networks*, (pp. 586–591).

Rosenberg, D. S. (2017). Bagging and Random Forests.

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*, 399–418.

Sauer, R. D. (1998). The economics of wagering markets. *Journal of economic Literature, 36*, 2021–2064.

Sigler, L. E. (2002). *Fibonacci's Liber Abaci: A Translation into Modern English of Leonardo Pisano's Book of Calculation.*

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

UEFA. (2021, 6). Abolition of the away goals rule in all UEFA club competitions. *Abolition of the away goals rule in all UEFA club competitions*. Retrieved from https://www.uefa.com/returntoplay/news/026a-1298aeb73a7a-5b64cb68d920-1000–abolition-of-away-goals-rule-in-all-uefa-club-competitions/

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/

Vlastakis, N., Dotsis, G., & Markellos, R. N. (2006). Beating the odds: Arbitrage and wining strategies in the football betting market. *Universidad Complutense, Madrid, Spain*.

Xiong, S., Yang, L., Zin, N. A., & Iida, H. (2016). Mathematical model of ranking accuracy and popularity promotion. *2016 3rd international conference on systems and informatics (ICSAI)*, (pp. 350–357).

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

Appendices

Appendix A: Variable Definitions

Table A-1: Definitions of all (numeric) variables.

Variable	Definition
Attendance	Number of people who attended the game at the stadium. https://www.worldfootball.net
Odds B365 Home Team Winning	Bet365 home win odds. www.football-data.co.uk/data.php
Odds B365 Draw	Bet365 home draw odds. www.football-data.co.uk/data.php
Odds B365 Away Team Winning	Bet365 home lose odds. www.football-data.co.uk/data.php
Rank Home Team	The pre-game ranking in the league of the home playing team.
Rank Away Team	The pre-game ranking in the league of the away playing team.
Attack Strength Home Team	Home playing team's average number of goals, divided by the league's average number of goals.
Attack Strength Away Team	Away playing team's average number of goals, divided by the league's average number of goals.
Defence Weakness Home Team	Home playing team's average number of goals conceded divided by the league's average number of goals conceded.
Defence Weakness Away Team	Away playing team's average number of goals conceded divided by the league's average number of goals conceded.
Average Points Home Team	The average number of points achieved by the home playing team in a season. The number of points is divided by the number of matches played.
Average Points Away Team	The average number of points achieved by the away playing team in a season. The number of points is divided by the number of matches played.
Average Losing Points Home Team	The average number of losing points achieved by the home playing team in a season. The number of losing points is the maximum possible number of points minus the actual number of points achieved by a team.
Average Losing Points Away Team	The average number of losing points achieved by the away playing team in a season. The number of losing points is the maximum possible number of points minus the actual number of points achieved by a team.
Average Goal Difference Home Team	The average goal difference for the home playing team in a season. The goal difference is divided by the number of matches played.
Average Goal Difference Away Team	The average goal difference for the away playing team in a season. The goal difference is divided by the number of matches played.
Home Wins Ratio Home Team	The ratio of home wins for the home playing team in a season.
Away Wins Ratio Away Team	The ratio of away wins for the away playing team in a season.
Home Draws Ratio Home Team	The ratio of home draws for the home playing team in a season.
Away Draws Ratio Away Team	The ratio of away draws for the away playing team in a season.
Form Home Team	The mean of number of points achieved in the previous five Premier League matches multiplied by the mean of the Elo rating of the opponents for the home playing team. For the first five match rounds in a season the means of the matches played till that match rounds are used.
Form Away Team	The mean of number of points achieved in the previous five Premier League matches multiplied by the mean of the Elo rating of the opponents for the away playing team. For the first five match rounds in a season the means of the matches played till that match rounds are used.
Elo Rating Home Team	Elo rating based on the Elo rating system for the home playing team. http://clubelo.com/ENG
Elo Rating Away Team	Elo rating based on the Elo rating system for the away playing team. http://clubelo.com/ENG
Home Field Advantage	A score to indicate how much the home team benefits from playing at home compared to the away team at a specific date. http://clubelo.com/ENG

Probability Home Team Winning (Elo)	Probability of a home win calculated by equation (4).
Probability Draw (Elo)	Probability of a draw calculated by equation (3)
Probability Away Team Winning (Elo)	Probability of an away win calculated by equation (5).
Win Ratio Home Team (H2H)	The head-to-head win ratio for the home playing team. https://www.soccerbase.com/
Win Ratio Away Team (H2H)	The head-to-head win ratio for the away playing team. https://www.soccerbase.com/
Draw Ratio (H2H)	The head-to-head draw ratio for both the home and away team. https://www.soccerbase.com/

Appendix B: Graphs

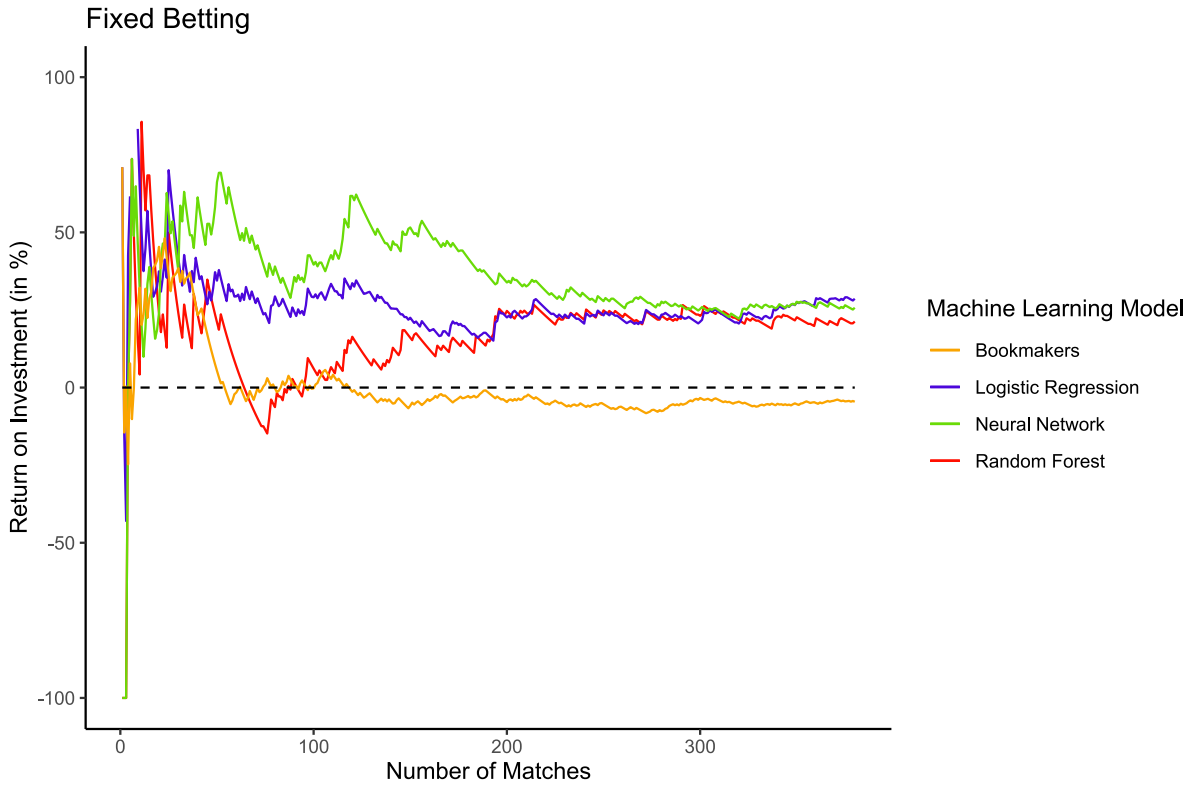

Figure B-1: Development of the return on investment for the second betting strategy (bet on misprised teams) with the fixed betting technique.

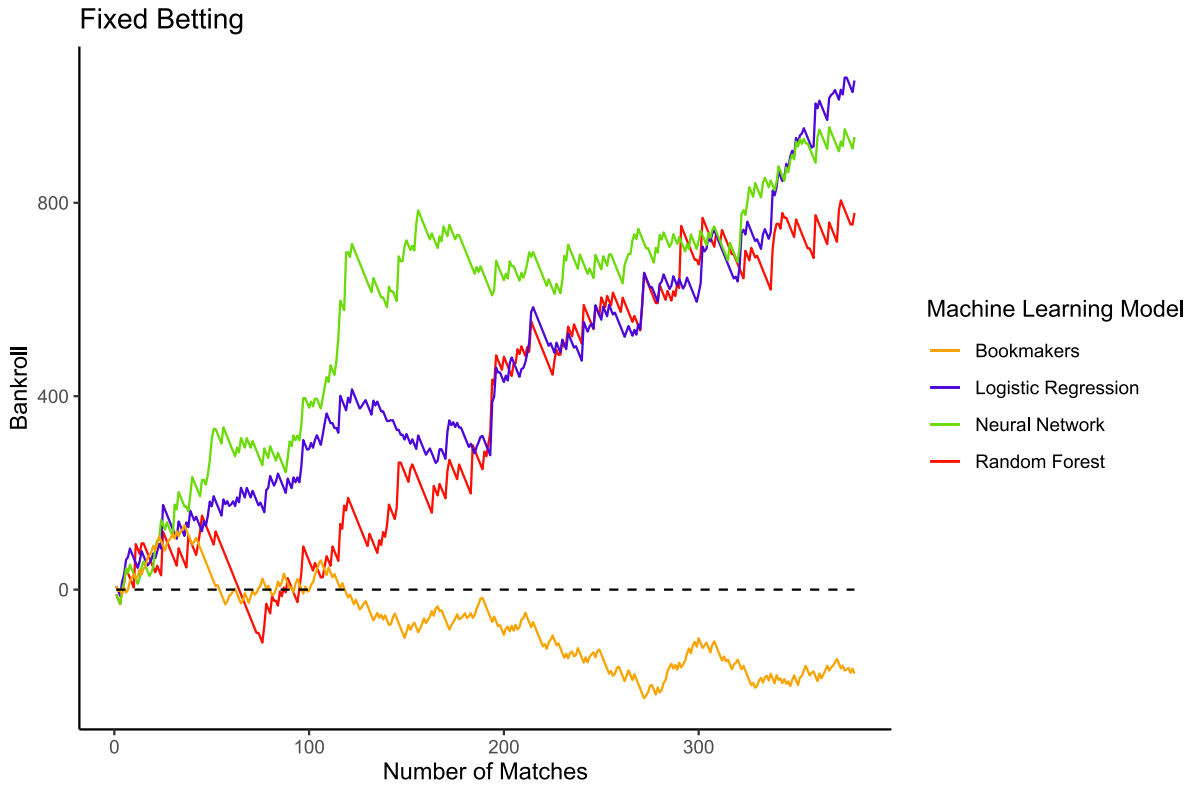

Figure B-2: Development of the bankroll for the second betting (bet on misprised teams) strategy with the fixed betting technique.

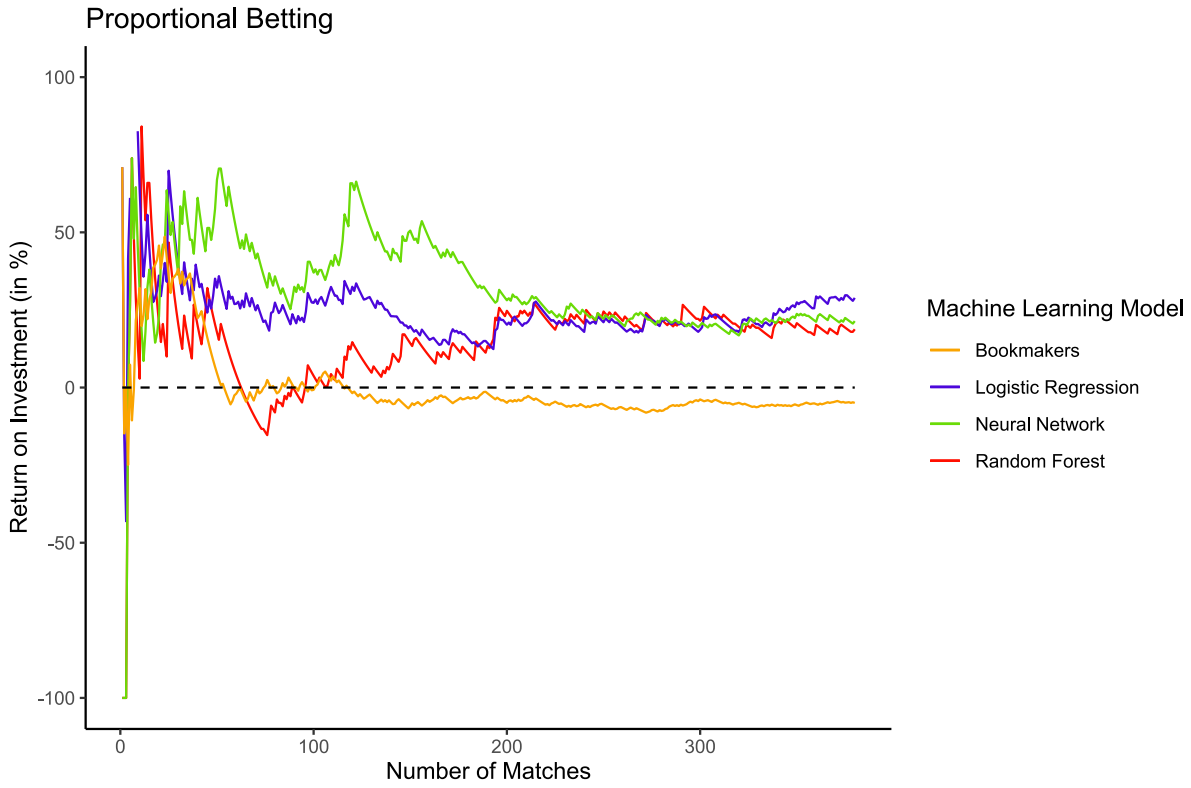

Figure B-3: Development of the return on investment for the second betting strategy (bet on misprised teams) with the proportional betting technique.

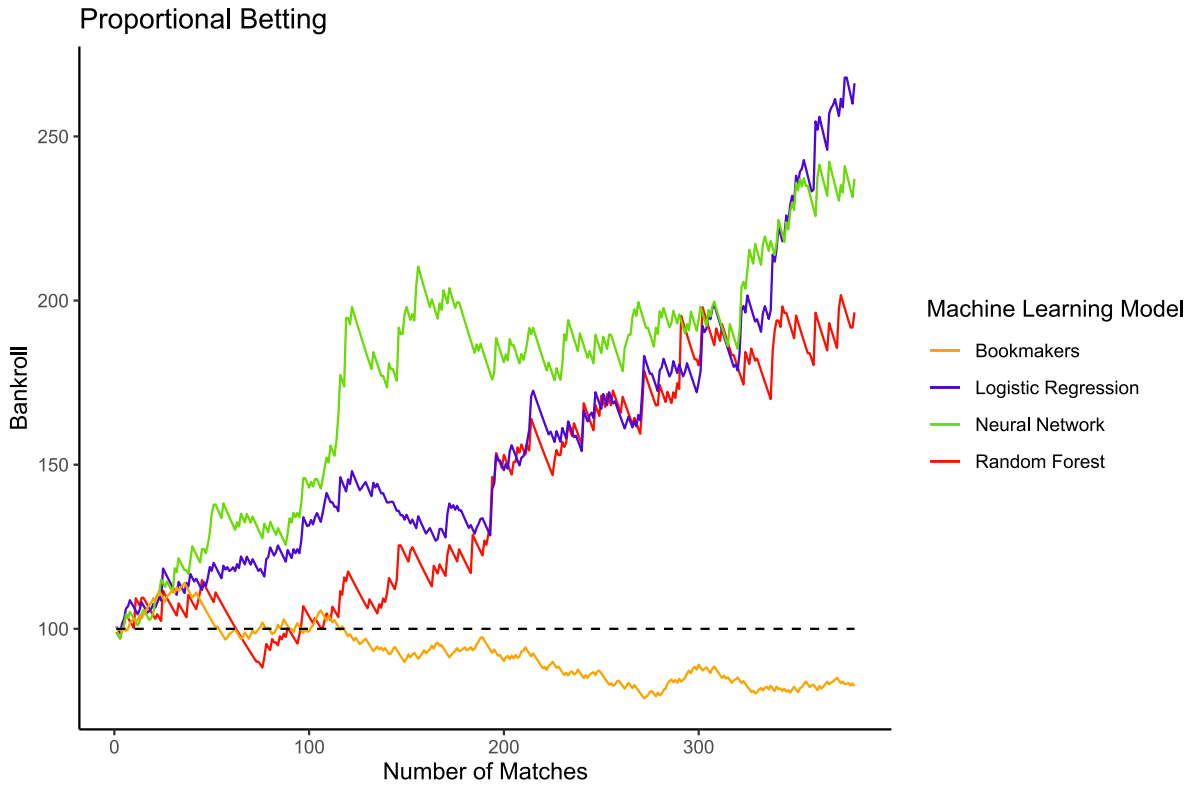

Figure B-4: Development of the bankroll for the second betting (bet on misprised teams) strategy with the proportional betting technique.

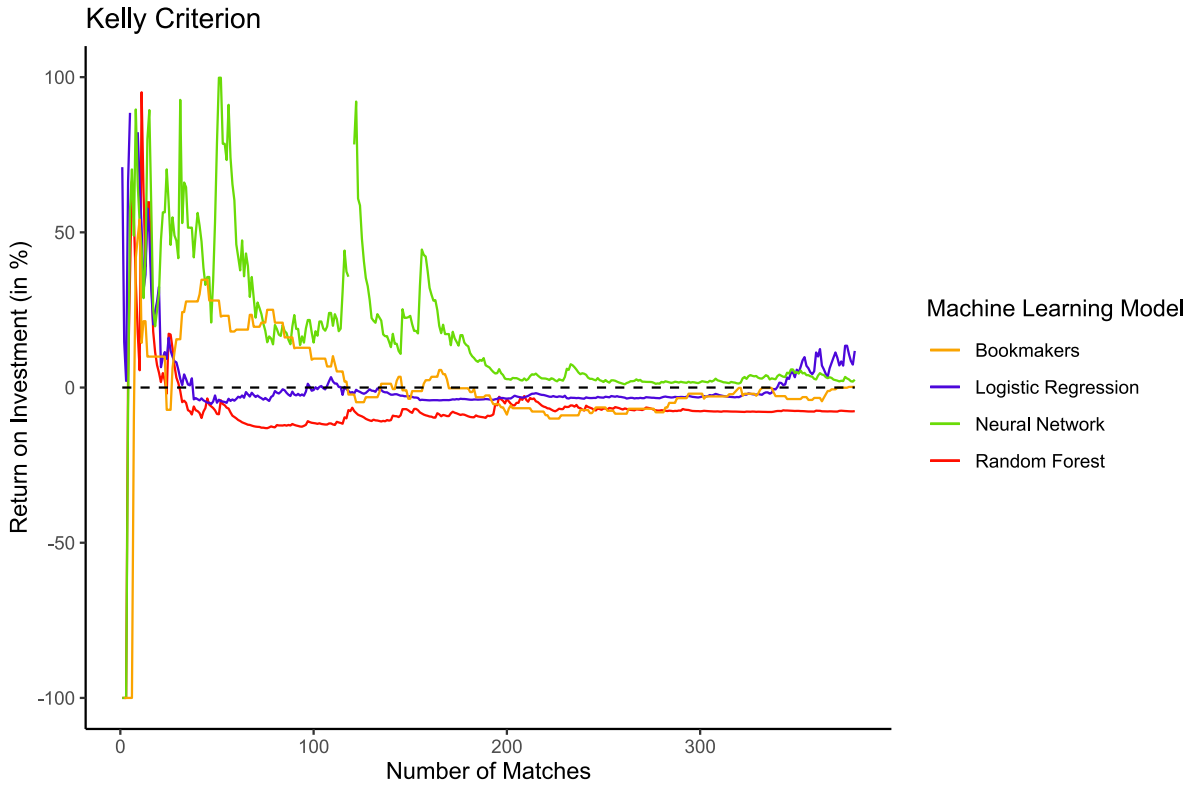

Figure B-5: Development of the return on investment for the second betting strategy (bet on misprised teams) with the Kelly criterion technique.

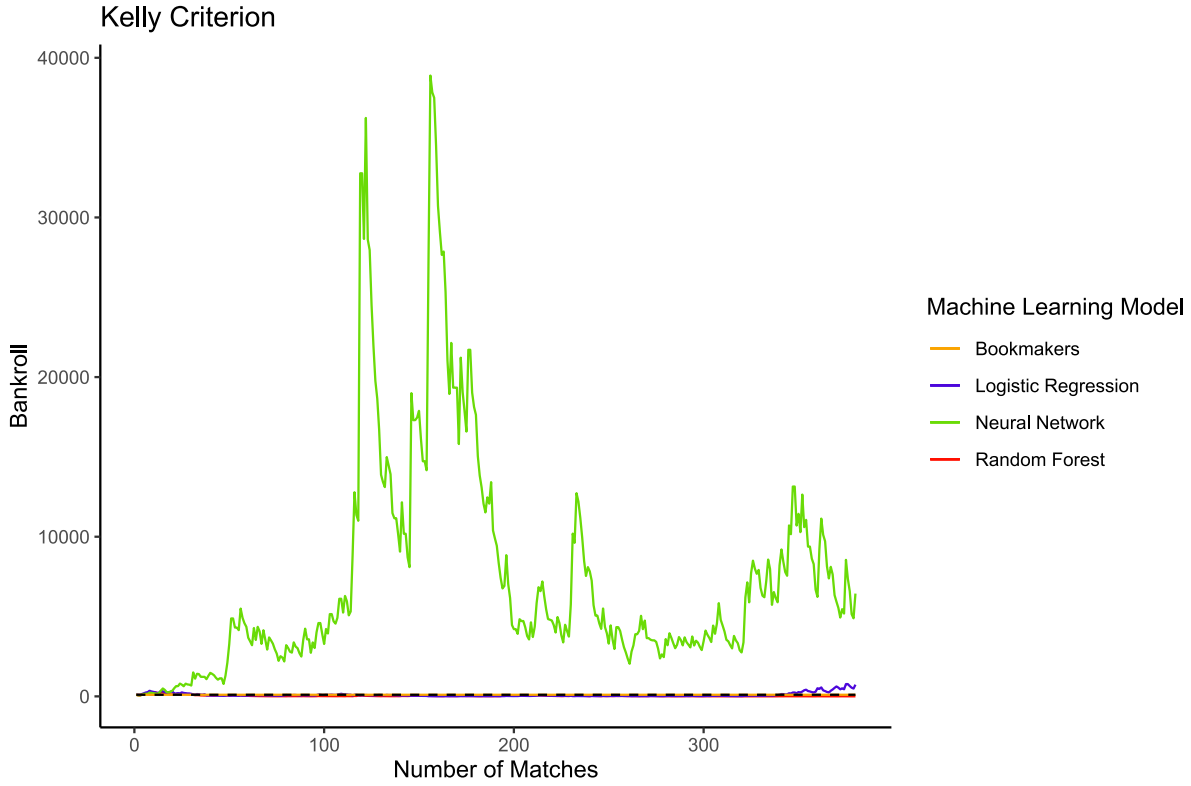

Figure B-6: Development of the bankroll for the second betting (bet on misprised teams) strategy with the Kelly criterion technique.

Figure B-7: Development of the return on investment for the second betting strategy (bet on misprised teams) with the fixed expected return technique.

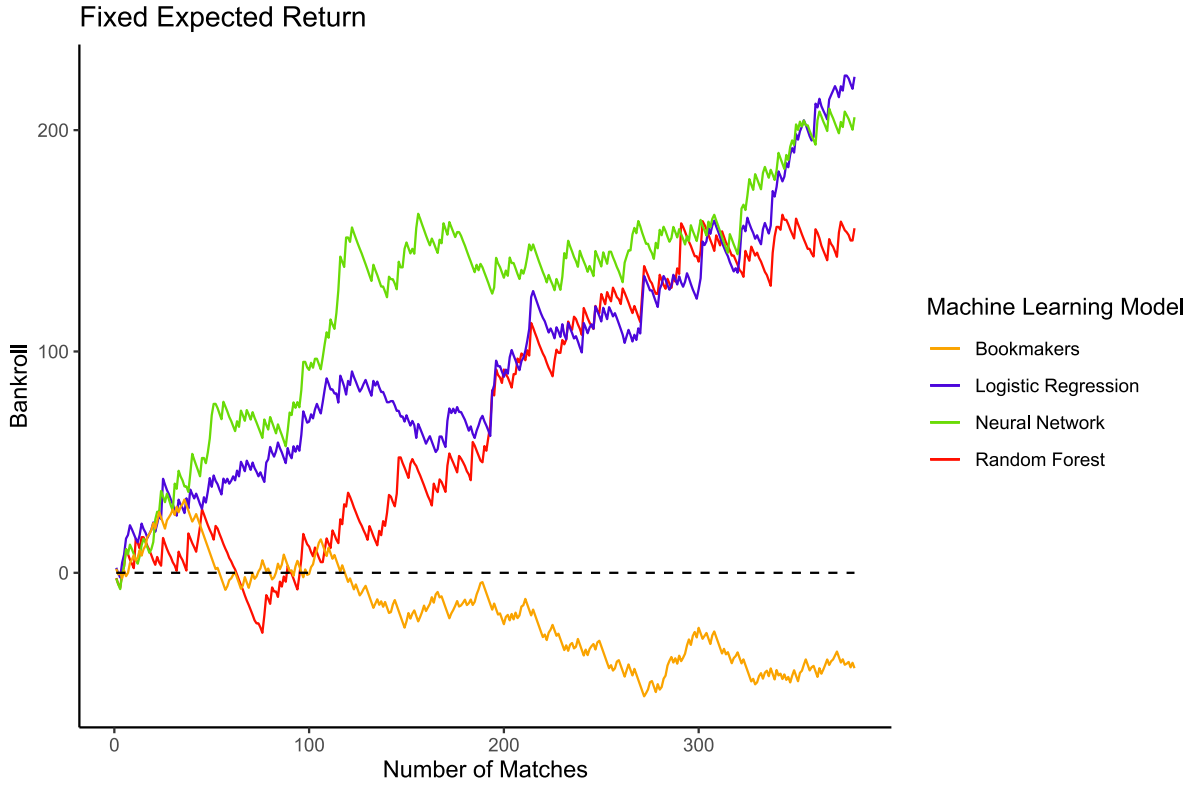

Figure B-8: Development of the bankroll for the second betting (bet on misprised teams) strategy with the fixed expected return technique.

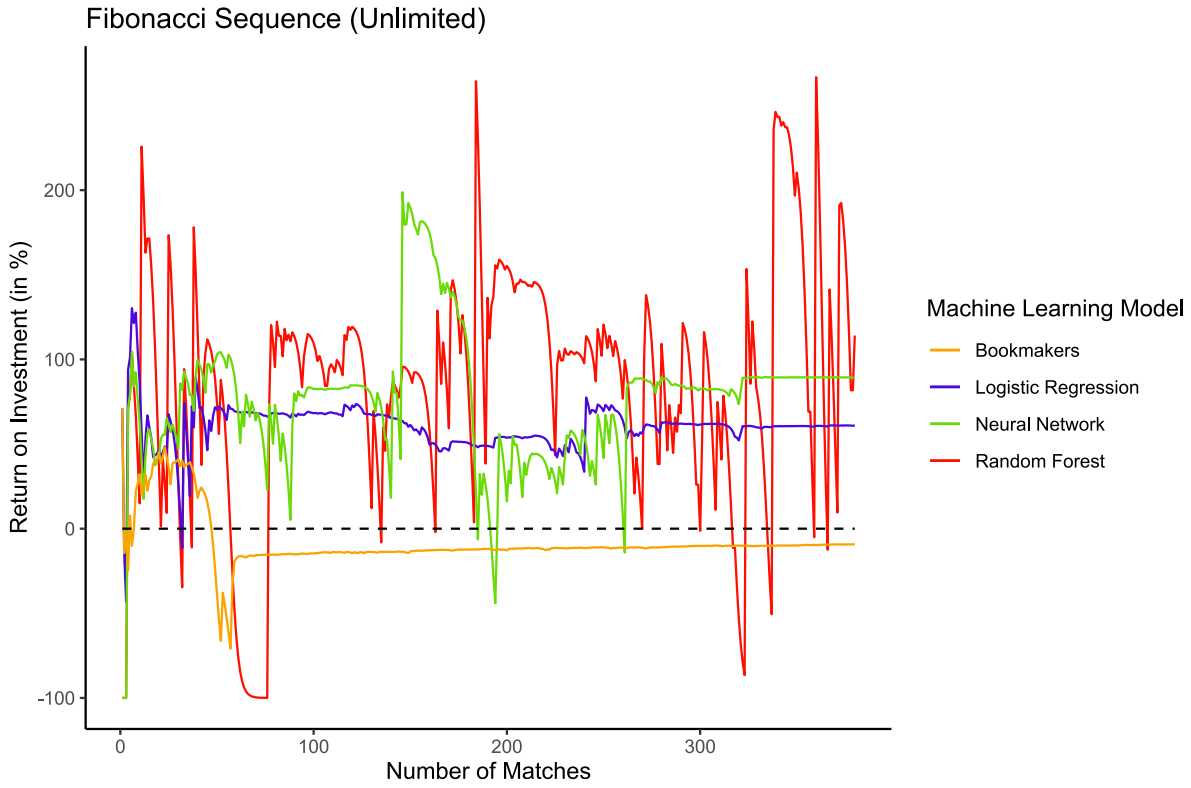

Figure B-9: Development of the return on investment for the second betting strategy (bet on misprised teams) with the Fibonacci sequence technique with no limit.

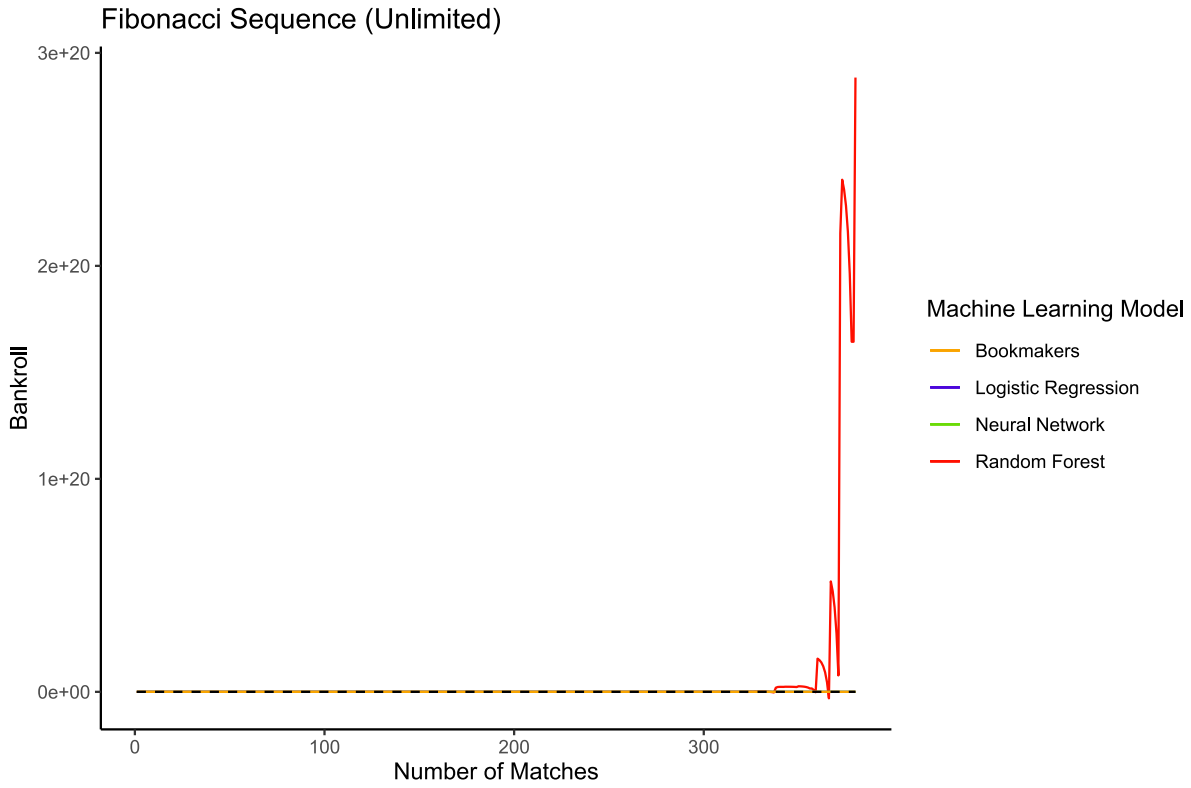

Figure B-10: Development of the bankroll for the second betting strategy (bet on misprised teams) with the Fibonacci sequence technique with no limit.

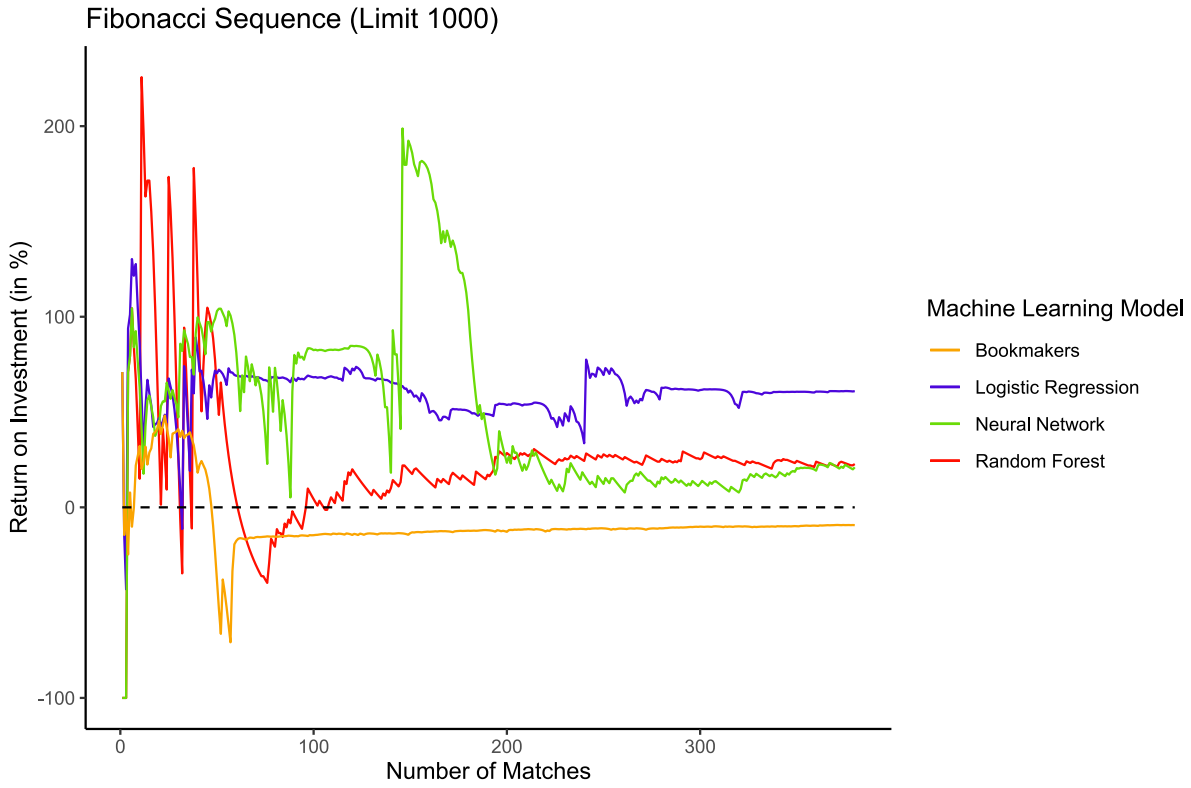

Figure B-11: Development of the return on investment for the second betting strategy (bet on misprised teams) with the Fibonacci sequence technique with a limit of 1.000 units.

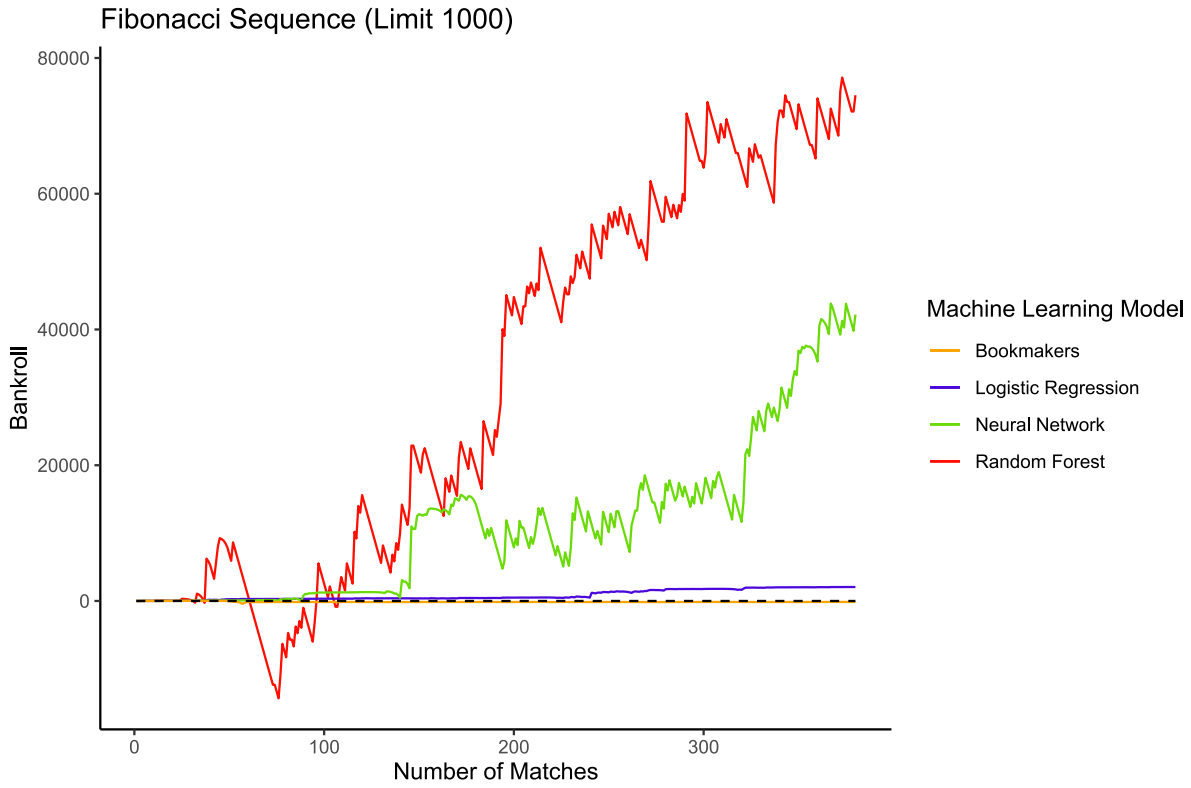

Figure B-12: Development of the bankroll for the second betting strategy (bet on misprised teams) with the Fibonacci sequence technique with a limit of 1.000 units.

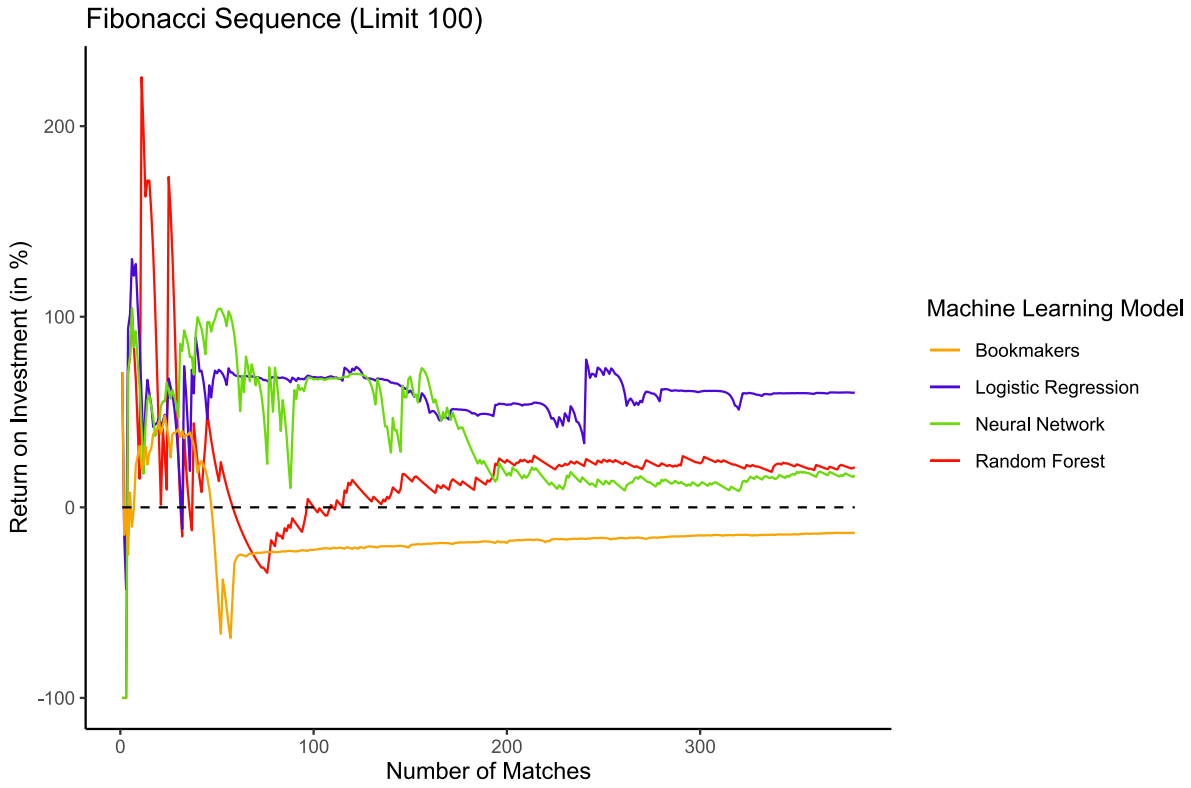

Figure B-13: Development of the return on investment for the second betting strategy (bet on misprised teams) with the Fibonacci sequence technique with a limit of 100 units.

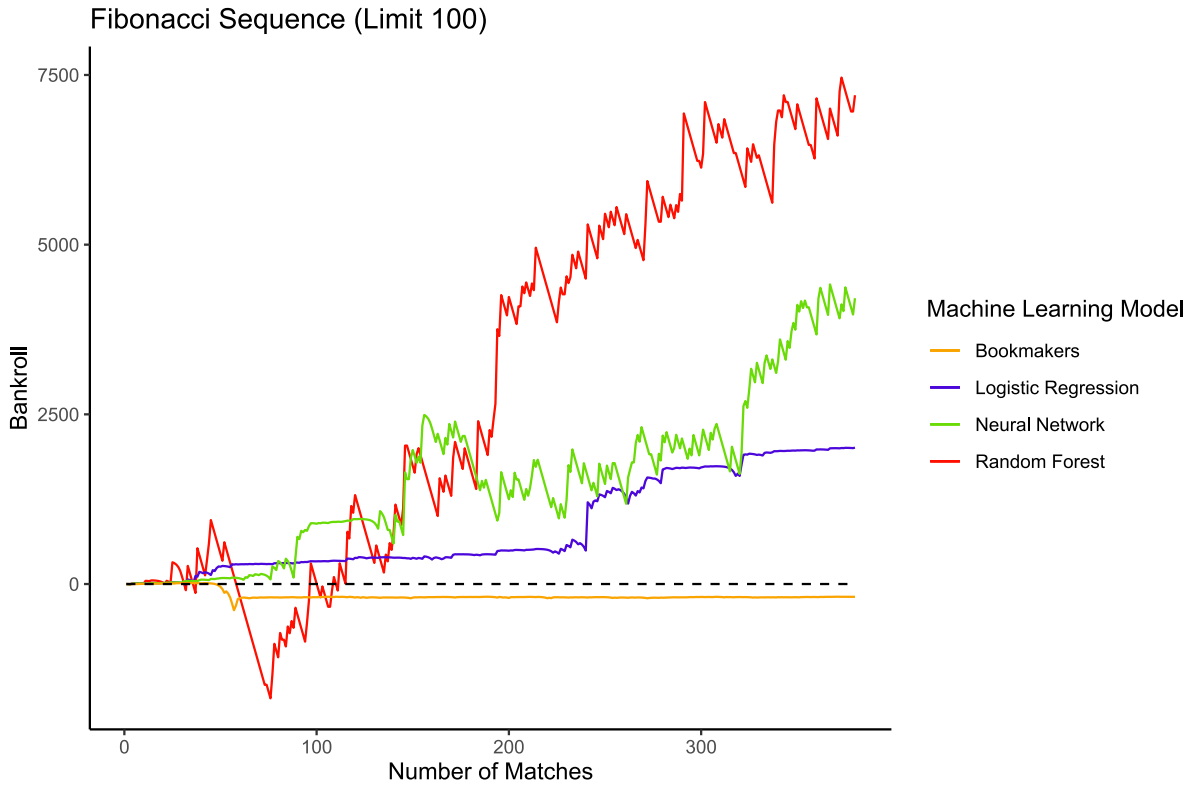

Figure B-14: Development of the bankroll for the second betting strategy (bet on misprised teams) with the Fibonacci sequence technique with a limit of 100 units.

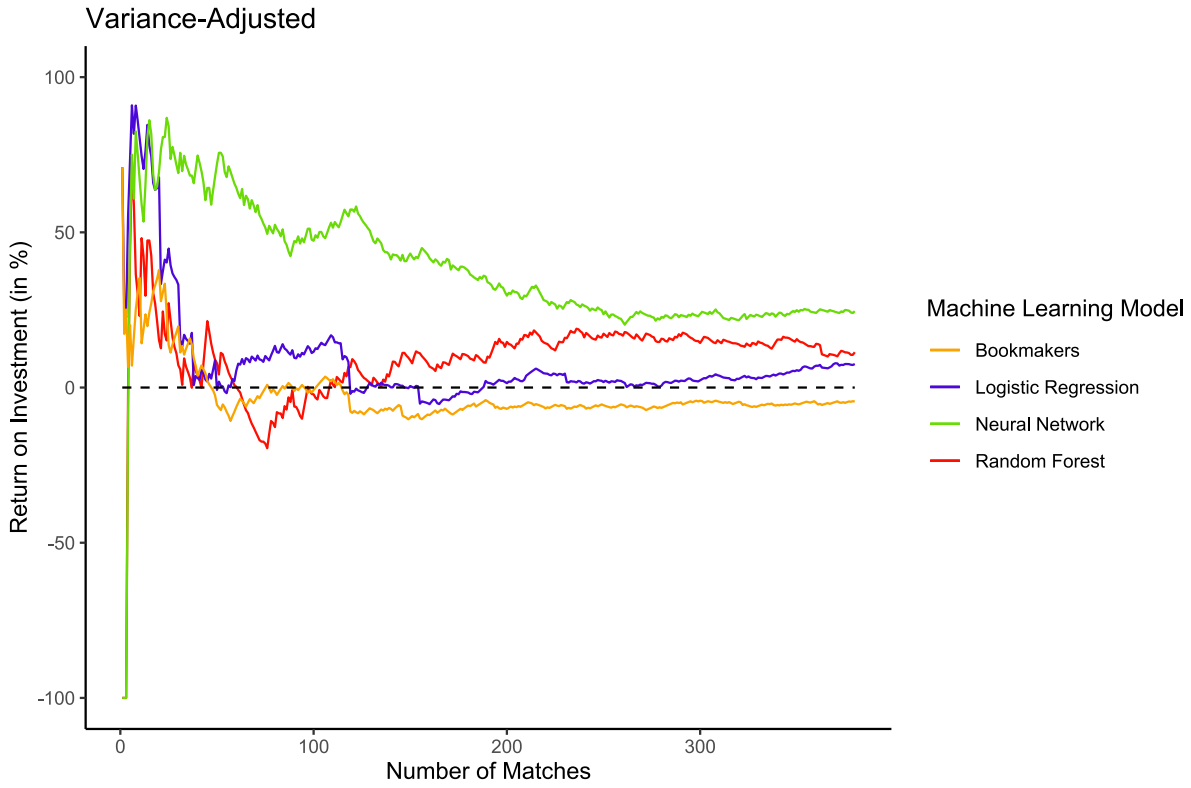

Figure B-15: Development of the return on investment for the second betting strategy (bet on misprised teams) with the variance-adjusted technique.

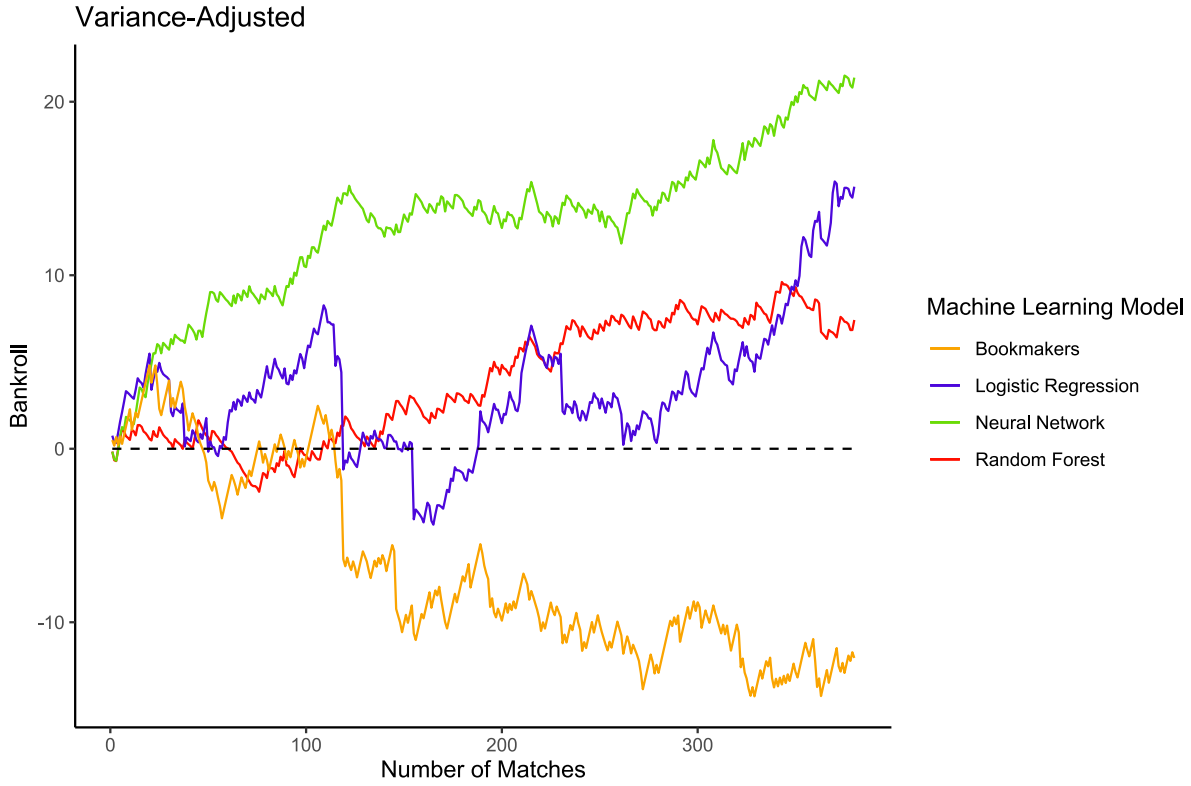

Figure B-16: Development of the bankroll for the second betting strategy (bet on misprised teams) with the variance-adjusted technique.