KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-019-00610-4

TECHNICAL CONTRIBUTION

Machine-Learning-Based Statistical Arbitrage Football Betting

Julian Knoll[1] · **Johannes Stübinger**[2]

Received: 11 March 2019 / Accepted: 6 July 2019

Abstract
Across countries and continents, football (soccer) has drawn increasingly more attention over the last decades and developed into a huge commercial complex. Consequently, the market of bookmakers providing the possibility to bet on the result of football matches grew rapidly, especially with the appearance of the internet. With a high number of games every week in multiple countries, football league matches hold enormous potential for generating profits over time with the use of advanced betting strategies. In this paper, we use machine learning for predicting the outcome of football league matches by exploiting data about match characteristics. Based on insights from the field of statistical arbitrage stock market trading, we show that one could generate meaningful profits over time by betting accordingly. A simulation study analyzing the matches of the five top European football leagues from season 2013/14 to 2017/18 presented economically and statistically significant returns achieved by exploiting large data sets with modern machine learning algorithms. In contrast to these modern algorithms, the break-even point could not be reached with an ordinary linear regression approach or simple betting strategies, e.g. always betting on the home team.

Keywords Football · Betting strategy · Machine learning · Statistical arbitrage · Sports forecasting

1 Introduction

For football fans, the outcome of the favorite team's next match is often a huge topic of discussion. As in many sports, the final result of the game is the only metric which ultimately determines a team's level of success. Due to the significant impact of the match result, many people attempt to predict the outcome of football games to earn not only admiration from other fans, but also money by betting on the winning team.

This paper presents a methodology for predicting the outcome of football matches by employing machine learning approaches. In this context, different approaches were compared to uncover whether higher complexity in the algorithm pays off in the form of a more precise forecast. The estimated outcomes were backtested based on the betting odds of one of the world's leading online sports betting providers and resulted in continuously positive returns. For this purpose, methods of statistical arbitrage stock market trading served as a framework for the backtest.

We make the following main contributions to the literature. First, we developed a strategy for betting on football league games, incorporating methods from both statistical arbitrage trading and machine learning. Second, we challenged our strategies based on real betting odds of an online bookmaker, which revealed positive returns over time. Third, we observed within a backtest very favorable risk-return characteristics as well as a higher performance of risk-averse strategies in contrast to risk-taking ones. Fourth, we analyzed in detail the influence thresholds make on the results and in which cases the machine learning approaches could succeed.

The remainder of this work is structured as follows. Section 2 contains an overview of the related work. After describing the data and simulation study in Sect. 3, we discuss the results in Sect. 5. Finally, we provide concluding remarks and an outlook on our future work in Sect. 6. In this article, "football" refers to the popular sport association

✉ Julian Knoll
julian.knoll@fom-net.de

Johannes Stübinger
johannes.stuebinger@fau.de

[1] Hochschule für Oekonomie and Management, Nuremberg, Germany

[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Nuremberg, Germany

Published online: 12 July 2019

football (soccer), not to other sports like American football or rugby.

2 Related Work

2.1 Literature on Sports Bets

2.1.1 Betting and Financial Markets

There are some publications regarding the efficiency betting exchange markets. For example, Gil et al. [22] examined the market during the 2002 FIFA World Cup and noted that their observations provide only mixed support for the hypothesis of efficient markets. In contrast, Croxson and Reade [11] conducted a similar study and found fast and full adjustment of prices. Despite an immediate increase of the prices after a goal, these prices still remained higher 10 to 15 minutes afterwards. Forrest and Simmons [17] analyzed the betting market of the top tier of Spanish football. They observed that the relative number of fans of each club in a match appears to have an influence on betting odds, with supporters of more popular teams offered more favorable conditions. Franck et al. published an inter-market study comparing the forecasting accuracy of bookmakers to a major betting exchange [18] and researched the inter-market arbitrage in betting [19]. They claimed that combining the data from bookmakers and from the bet exchange market yields a guaranteed positive return. Moreover, empirical studies and meta-studies about the accuracy of sports experts and the bet exchange market [50, 53] predicting the outcome of a sporting event were published. Choi and Hui [10] analyzed the reaction to unanticipated events using the in-play football betting market. They found that most market participants under-react to new events but over-react to events that are highly surprising.

Furthermore, some articles researched the impact of match results of publicly traded sports teams on the corresponding stock prices. Palomino et al. [41] investigated how football team stocks react regarding the outcome of a match. They observed that abnormal returns for winning teams do not reflect rational expectations, but are high due to overreactions induced by investor sentiment. Levitt [34] compared sports betting markets and financial markets for NFL football teams and found that, due to the ability of bookmakers to set prices, they make greater profits than they could generate by acting like traditional brokers and attempted to balance out supply and demand. Bernile and Lyandres [3] examined whether investors' biased ex-ante beliefs regarding outcomes of a future event can explain inefficient stock markets. They examined stocks of publicly traded European football clubs around important matches and found that an investor sentiment is attributable in part to a systematic bias in the investors' ex-ante expectations.

2.1.2 Prediction of The Outcomes of Sporting Events

Some publications about the prediction of the outcome of football matches refer specifically to major sporting events, such as FIFA World Cup or UEFA Euro Cup. Stefani [51] had already introduced a least squares betting approach in 1980 and applied it to data about the FIFA World Cup 1976. Archontakis and Osborne [1] formulated a strategy, which employed the Fibonacci sequence to generate bets. Conducting a simulation on FIFA World Cup finals, they claimed it is possible to earn economic profits through this method with fairly large risk. Luckner et al. [38] published an empirical study that compared the accuracy of a prediction market of the FIFA World Cup 2006 to predictions derived from the FIFA world ranking. They found that prediction markets for the FIFA World Cup outperform predictions based on the FIFA world ranking in terms of forecast accuracy. Groll et al. [24–26] worked on covariate-based prediction of major international football tournaments such as European championships and FIFA World Cups, starting from the EURO 2012. For their prediction, they also use modern techniques such as regularized regressions and random forests. Zeileis et al. [67] introduced a probabilistic prediction for the 2018 FIFA World Cup based on the bookmaker consensus model to find the winner of the FIFA World Cup. Based on a similar strategy, they predicted the winner of the UEFA Euro Cup 2 years before [66].

In addition to predicting the FIFA World Cup, Stefani [51] used the least squares betting approach to forecast results of other sports such as American football and basketball as well. Moreover, Lisi and Zanella [36] investigated a betting strategy based on about 500 tennis matches resulting in a cumulative return of 16%.

2.1.3 Prediction of Football League Match Results

This paper aims to predict the outcome of football league matches based on characteristics and skills of the football players involved. In this context, we only found a few publications in the periphery of this topic:

– Maher [39] published fundamental research regarding predicting football league matches in 1982. He investigated a Poisson model based on parameters representing the teams' inherent attacking and defensive strengths and found the most appropriate model from a hierarchy of models. The observed and expected frequencies of scores are compared afterwards and goodness-of-fit tests reveal that despite some small systematic differences, an inde-

pendent Poisson model can describe football scores with a reasonable accuracy.
- Dixon and Coles [12] analyzed English league and cup football data from 1992 to 1995. They developed a model trying to exploit potential inefficiencies in the football betting market using bookmakers' odds from 1995 to 1996. Their approach is based on a Poisson regression model enhanced by the data structure and the dynamic nature of performances of the involved football teams. The model had a positive return when used as a basis for a betting strategy.
- Rue and Salvesen [46] applied a Bayesian dynamic generalized linear model to extract time dependent skills of the teams of the English Premier League and the Spanish Primera Division. Based on the match results of former matches, they used an algorithm to find the parameters of their model and to predict the next football matches. For a total of 3892 football matches between 1993 and 1997, they reveal a final cumulative return of 40% for the English Premier League and 54% for the Spanish Primera Division.
- Godin et al. [23] described another approach for predicting Premier League football matches. They analyzed how to incorporate Twitter Microposts in which users made a guess how a football match would end up. Therefore, this information was extracted from unstructured text by a parsing algorithm. They predicted about 200 match results in 2013/14 and claimed to be able to realize a (theoretical) profit of 30%.
- Koopman and Lit [32] proposed a model for football match forecasting based on a bivariate Poisson distribution with intensity coefficients which change stochastically over time. It is based on state space and importance sampling methods which are computationally efficient. The out-of-sample performance of their methodology was applied to the match outcomes from the 2010–2011 and 2011–2012 seasons of the English football Premier League. They observed that their statistical modelling framework can produce a significant positive return over the bookmaker's odds.
- Tax and Joustra [58] predicted match results of the Dutch Eredivisie between 2000 and 2013. Their data consisted mainly of the results of the former matches. In addition, they incorporated data about whether the team played in a lower league the season before, whether a new coach was hired, or whether the top scorer was injured. Based on these data, different machine learning algorithms were analyzed regarding their accuracy (e.g., Naive Bayes, Neural Networks, or Decision trees).
- Schauberger et al. [49] intended to find the on-field variables connected to the sportive success of football teams. They propose an extended Bradley–Terry model for football matches which takes on-field covariates into account and use penalty terms to reduce the complexity of the model. Their model identified the running distance to be the on-field covariate with the strongest to the match outcome.
- Stübinger and Knoll [56] predicted the outcome of selected football matches based on the corresponding team characteristics. They exploited the obtained information using a risk-averse betting strategy. For a total of 8082 football matches between 2013 and 2018, the authors found economically and statistically significant returns.
- Egidi et al. [14] developed a hierarchical Bayesian Poisson model with the scoring rates of the teams being represented by convex combinations of parameters estimated from historical data and betting odds. They applied their approach to a 9-year dataset of four European football leagues to predict match outcomes for the tenth season.

In conclusion, no published article exists which examines the topic this paper focuses on. Some articles aim to connect the field of betting with financial markets ([10, 19, 22]) while others forecast sporting events ([24, 36, 51, 67]). Currently, there is no study about predicting and betting on football matches over several years for the big five football leagues (England, France, Germany, Italy, Spain) based on the corresponding match characteristics.

2.2 Literature on Statistical Arbitrage

In the mid-eighties, a group of mathematicians and physicists at Morgan Stanley developed statistical arbitrage as a trading strategy. As a long-term trading opportunity, statistical arbitrage exploits persistent capital market anomalies to create economically and statistically significant profits over time. The corresponding data-driven trading recommendations are based on interdisciplinary methods, such as state-of-the-art models from the fields of mathematics, physics, computer science and operations research.

In recent years, interest in statistical arbitrage trading has increased noticeably in the academic community. Research work in this field examines either theoretical principles or empirical applications. The key representatives are Gatev et al. [21], Avellaneda et al. [2], Bertram [4], and Liu et al. [37]. To date, there is only one study presenting a statistical arbitrage strategy in the field of sports betting [56]. Since this article serves as an extended version, it provides additional analysis and in depth insights, e.g. for which outcome the different approaches succeed and regarding the influence thresholds make on the results.

Table 1 Summary of the match characteristics data from season 2013/14 to 2017/18

	Home team				Away team			
	Min	Median	Max	Mean	Min	Median	Max	Mean
General game								
Ball possession	15.5%	51.6%	83.4%	51.3%	16.6%	48.4%	84.5%	48.7%
Duel quota[1]	28.4%	50.4%	66.7%	50.3%	33.0%	49.5%	70.5%	49.5%
Air duel quota	11.5%	50.0%	100%	50.7%	0.0%	50.0%	88.5%	49.3%
Intercepted balls	0	15	47	15.47	0	15	43	15.78
Number of offsides	0	2	14	2.38	0	2	14	2.15
Number of corners	0	5	20	5.63	0	4	18	4.43
Pass								
Number of passes	135	433	1015	447.79	161	412	1078	426.25
Long passes proportion	2.3%	14.7%	40.7%	15.2%	2.9%	15.4%	42.7%	16.0%
Pass quota[2]	41.9%	78.7%	94.1%	78.0%	41.0%	77.5%	93.1%	76.7%
Pass quota (opposing half)	31.7%	70.5%	92.8%	70.2%	32.8%	68.8%	92.4%	68.5%
Number of crosses	1	20	81	21.36	1	16	55	16.83
Cross quota[3]	0.0%	23.1%	75.0%	23.6%	0.0%	22.2%	100%	22.7%
Defense and discipline								
Number of tackles	3	18	48	18.80	3	19	45	19.07
Tackle quota[4]	22.2%	75.0%	100%	74.3%	0.0%	74.1%	100%	73.6%
Clarifying actions[5]	1	20	68	21.52	2	24	89	25.55
Number of fouls	2	13	33	13.41	0	14	32	13.80
Number of sending-offs	0	0	2	0.10	0	0	3	0.13
Attack								
Goals	0	1	10	1.55	0	1	9	1.17
Shots	1	13	43	13.94	0	11	35	11.22
Shots at the goal	0	5	16	4.85	0	4	15	3.89
Shots inside the box	0	8	26	8.20	0	6	23	6.38
Shots outside the box	0	5	23	5.74	0	5	19	4.84
Shooting accuracy	0.0%	45.5%	100%	45.3%	0.0%	44.4%	100%	44.9%

[1] Number of successful duels divided by the total number of duels

[2] Number of successful passes divided by the total number of passes

[3] Number of successful crosses divided by the total number of crosses

[4] Number of successful tackles divided by the total number of tackles

[5] Number of successful defense actions

3 Simulation Study

3.1 Data Sources

For our empirical application, we collected football match data from the five top European leagues, i.e., the Premier League, Ligue 1, Bundesliga, Serie A, and Primera Division from season 2013/14 to 2017/18. This data set serves as a true acid test for any back-testing study because analyst coverage and investor scrutiny is especially high for these large capitalized leagues. If a match result was subsequently changed by an arbitrating body for other reasons, the original result was still used. In the following, the data set used within our simulation study is described in more detail.

3.1.1 Match Characteristics

Table 1 presents a summary (minimum, median, maximum, and mean) of the match characteristics of all 8082 football matches analyzed. The match characteristics were downloaded from Sportal[1]. The data set contains information about the general game, pass behavior, defense and disciplinary measures, and attack capacities for both the home and the away team. The table provides a good illustration of the well-known home advantage. The average values of most characteristics show higher values for the home team. Only attributes characterizing defensive behavior tend to

[1] We thank https://www.sportal.de/ for providing the data.

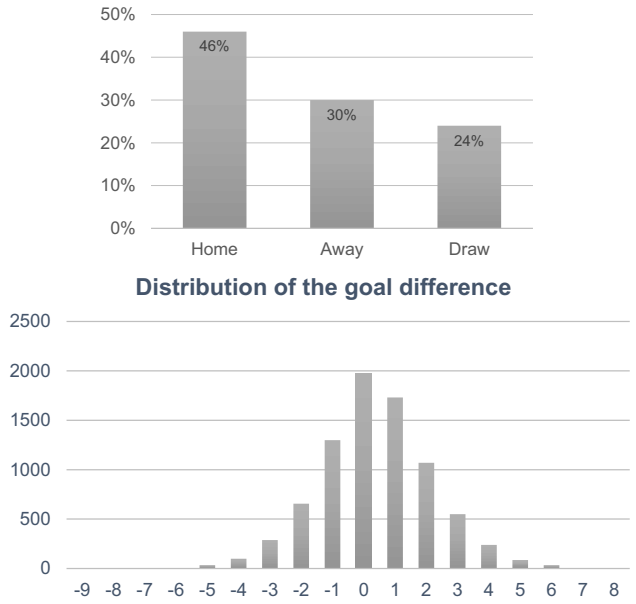

Fig. 1 Properties of the match results from season 2013/14 to 2017/18

be greater for the away teams, e.g., "number of intercepted balls", "number of clarifying actions", "number of fouls", or "number of sending-offs".

3.1.2 Match Results

The goal statistics for the analyzed football matches are presented in Fig. 1 in more detail. As mentioned above, the main focus of this research paper is to predict which team would win a specific football league match. Therefore, the difference between the number of goals of the home team and the number of goals of the away team plays a decisive role.

In total, 3723 home team wins (46%), 1967 draws (24%), and 2392 away team wins (30%) were observed. This illustrates the home advantage also found in the match characteristics. This fact is well in line with the finding that the distribution of the goal difference (home team goals minus away team goals) is asymmetric, which means that there are more matches with a positive goal difference (home teams wins) than with a negative goal difference (away team wins). Though, the most frequent outcome of the football match was no difference between home and away team (draw).

3.1.3 Betting Odds

To obtain a financial performance evaluation of our approach, we also collected the betting odds corresponding to the analyzed football matches from the online bookmaker

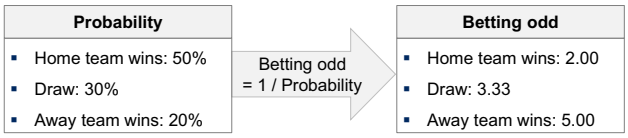

Fig. 2 Relation between betting odds and corresponding probabilities

Table 2 Summary of the betting odds from season 2013/14 to 2017/18

	Min	Median	Max	Mean
Home team wins	1.02	2.15	26	2.793
Draw	2.50	3.50	17	4.077
Away team wins	1.08	3.50	51	4.957

Bet365[2], one of the leading betting suppliers with about 23 million customers globally. The analysis is based on the very common decimal odds, which work as follows: A bet amount b is placed on a specific event and a given betting odd o. If the event occurs, the bet amount multiplied by the betting odd $b \cdot o$ is paid out. If the event does not occur, nothing is paid out. Thus, the relative return on b is always either $o - 1$ (for a successful bet) or -1 (for an unsuccessful bet), which means a complete loss of b.

Figure 2 depicts the relation between the decimal betting odds and the probabilities estimated by the bets broker, e.g. the odd of 2.0 for a home team win results in a estimated probability of 50 %. In this example, the three depicted probabilities sum up to exactly one since there are only three possible outcomes of a game (home win, draw, away win). With a setting like this, the bets broker would not make profit on the long term. Therefore, the bets broker subtracts a small proportion of every decimal odd. By doing so, the provider pays out less than expected based on the estimated probabilities and can consequently generate profits over time.

Table 2 shows a short summary of the collected betting odds. It is remarkable that a win for one of the teams leads to far more extreme betting odds than for a draw. Apart from this, it is hardly surprising that the bookmaker is aware of the aforementioned home advantage which can be seen by comparing the average odds of a home win with the average away win odds.

3.2 Software

The entire methodology and all relevant analyses were implemented in the statistical programming language `R`

[2] We thank https://www.football-data.co.uk/data.php for providing the data.

Table 3 R packages used in this paper for dependence modeling, data handling, and financial modeling

1. Dependence modeling	
randomForest	[35]
party	[28]
xgboost	[9]
e1071	[40]
2. Data handling	
dplyr	[62]
readr	[63]
readxl	[61]
texreg	[33]
xlsx	[13]
xts	[48]
zoo	[65]
3. Financial modeling	
PerformanceAnalytics	[42]
QRM	[43]
quantmod	[47]
sandwich	[64]
timeSeries	[45]
tseries	[59]
TTR	[60]

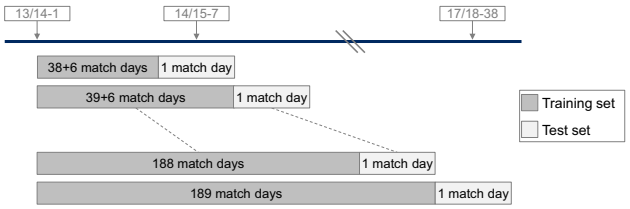

Fig. 3 Setup of the simulation study with a training period of at least one saison and a one match day prediction period

[44]. For computation-intensive calculations, we used the general-purpose programming language C++. Table 3 lists the additional packages for dependence modeling, data handling, and financial modeling.[3]

4 Methodology

4.1 Simulation Setup

The simulation study aimed (1) to predict football matches with the aid of data-driven methods and (2) to exploit the obtained information using a statistical arbitrage strategy.

[3] Please contact the authors if you are interested in the data and the code.

Specifically, different machine learning approaches were applied to forecast the goal difference between home team and away team (dependent variable y) based on the football match data observed up to that point (independent variables x). According to Jegadeesh and Titman [29], Gatev et al. [21], and Tax and Joustra [58], the data set was divided into overlapping study periods, each shifted by one match day. Figure 3 shows that each study period included a test set, which represents the considered match day, and a training set, which contains all previous match days. Consequently, the training set grew from 44 match days (one season plus the six first matches of the second season) to 189 match days (five seasons minus one match day). It is important that no information about future events was included in the data at any time during the prediction process.

4.1.1 Training Set

As mentioned above, the training set contains all information about the previous match days. Specifically, y_{train} describes the goal difference between home team and away team and x_{train} defines the corresponding football match data. The set of independent variables x_{train} includes the characteristics from the sections "general game", "pass", "defense and discipline" as well as "attack" (see Table 1). The variable "shots inside the box" was excluded from x_{train}, since it can be calculated by "shots" and "shots outside the box". Based on a similar thought three other variables were removed: The knowledge of "ball possession", "duel quota", and "air duel quota" of the home team makes the respective variable for the away team redundant. To model the relation of the dependent variable y_{train} as a function of the 39 independent variables x_{train}, different common machine learning approaches were employed:

– Random forest (RFO): an important representative of machine learning is the random forest, which is an ensemble learning method that works by building a variety of decision trees to output the class, i.e. the mean prediction of each tree. The approach corrects the habit of decision trees to overfit to the training set and decorrelates the single trees from the different (similar) bootstrap samples (compared to bagging). The tuning parameters are determined based on proposals from the literature, as we focused on our methodology. Of course, the optimization of these parameters would lead to a higher accuracy. For RFO, the number of trees is set to 500 and the number of variables randomly sampled as candidates considered for splitting to approximately 1/3 of the number of our variables. For further details about this approach, see Breimann [7] and Hastie et al. [27]. For R implementation, we use the package randomForest developed by [35].

- Boosting (BOO): boosting is a meta-algorithm for machine learning that merges several weak classifiers into a single strong classifier. Specifically, our gradient boosting approach reduces bias and variance. Again standard parameters are set, e.g., the maximum number of boost iterations is 400 and the learning rate is 0.05. Furthermore, the learning task and the corresponding learning objective is a linear regression with linear covariate effects. Please refer to Friedmann [20], Bühlmann and Hothorn [8], and Zhou [68] for more information. For `R` implementation, we use the package `xgboost` developed by [9].
- Support vector machine (SVM): support vector machine divides a set of objects into classes in such a way that as wide an area as possible remains free of objects around the class boundaries. The kernel trick is used in the case of non-linear separable data. In our manuscript, we used a kernel with a radial base. Steinwart and Christmann [52] described this approach in more detail. For `R` implementation, we use the package `e1071` developed by [40].

It should be noted that all of our machine learning approaches are regression models.

4.1.2 Test Set

For each study period, the aim was to predict the match results $\hat{y}_{\text{test}}$ of the considered match day. The corresponding match data were not used as independent variables x_{test}, since only information before the match starts was included to avoid any look-ahead bias. Therefore, x_{test} was determined based on the team characteristics of the last matches. For each home (away) team, the average characteristic value of each considered variable was calculated based on the past three home (away) matches—this procedure avoids any look-ahead bias and takes the home advantage into account[4]. Specifically, we determined the weighted median, a measure that is more robust against outliers than the weighted mean. It should be noted that home (away) teams had to possess at least team characteristics for three home (away) matches of the current season. Finally, the dependent variable was predicted based on the fitted models (RFO, BOO, SVM) and the team characteristics x_{test}.

This simulation study assumed that past team characteristics contain pieces of information that have a substantial effect on match results in the future. If our hypothesis holds and this relationship is reflected by the outlined models, we would be able to identify market inefficiencies in order to draw positive expected profits. The aim was to capture this circumstance with the following statistical arbitrage strategy:

- $\hat{y}_{\text{test}} > 2$ means the model predicted that the home team will win. Consequently, we bet 1 monetary unit on the home team, i.e. $b = 1$.
- $\hat{y}_{\text{test}} < -2$ means the model predicted that the away team will win. Consequently, we bet 1 monetary unit on the away team, i.e. $b = 1$.
- $-2 \leq \hat{y}_{\text{test}} \leq 2$ means the model does not provide a clear sign for "home team wins" or "away team wins". Consequently, we do not execute any bets, i.e. $b = 0$.

The trading thresholds of ± 2 were introduced because increasing deviations of the estimation from 0 depict higher chances of success. Since we are a conservative investor, the thresholds to ± 2 result from the objective of only betting on clear predictions. This parameter setting is consistent with the literature—the large-scale analyses of Bollinger [5] and Stübinger and Endres [55] show that trading thresholds of ± 2 are optimal in the context of financial bets. Another possible trading strategy would be to maximize the expected return [6, 24, 30, 32]. In general, one would choose the outcome with the highest expected return and only place the bet if the expected return is positive.

To assess the value-add of the trading strategies based on several machine learning approaches, they were benchmarked with a variant based on linear regression (LIR). This well-known method is a linear approach to model the relationship between dependent variables and one or more independent variables. The statistical properties of the resulting estimators can easily be determined.

Furthermore, three strategies based on the betting odds served as a benchmark: (1) Strategy BET bets 1 monetary unit per match on the outcome of the match result with the lowest odd. If two odds are identical, no action is taken. In this context, the lowest betting odd means always either "home team wins" or "away team wins" because "draw" never shows the lowest odd. (2) Strategy HOM bets 1 monetary unit per match on the outcome "home team wins". Other circumstances are not taken into account. (3) Strategy RAN randomly bets 1 monetary unit per match on the outcome "home team wins" or "away team wins".

5 Results

5.1 Statistical Analysis

The general results of the simulation study are shown in Table 4. While the rows are divided into different characteristics, the columns each reflect a different method or betting

[4] Without loss of generality, our model can also be used for matches without home advantage, e.g., FIFA World Cup and UEFA Euro Cup. In this case both teams would be neutral teams.

Table 4 Statistical characteristics for the strategies RFO, BOO, SVM, LIR, BET, HOM, and RAN from season 2013/14 to 2017/18

	RFO	BOO	SVM	LIR	BET	HOM	RAN
Prediction quality							
Accuracy	75.62%	70.95%	66.12%	63.60%	54.49%	46.10%	37.82%
RMSE	1.9986	2.2441	2.2899	2.2740	9.3875	9.7927	10.1494
MAD	1.5909	1.7964	1.8610	1.8631	9.2288	9.6182	9.9739
Betting details							
Number of bets	324	482	614	566	6586	6664	6664
Bet on home team	89.51%	79.67%	81.43%	71.20%	71.24%	100%	51.19%
Bet on away team	10.49%	20.33%	18.57%	28.80%	28.76%	0.00%	48.81%
Average payoff	1.0542	1.0217	1.0054	0.9957	0.9705	0.9676	0.9370
Predicted values							
Maximum	4.1788	4.8992	4.5194	4.2173	–	–	–
Minimum	−3.4892	−4.6230	−4.1254	−5.7978	–	–	–

Fig. 4 Comparison between observed and predicted goal difference per machine learning algorithm

RFO

Prediction \ Observation	≤ -5	-4	-3	-2	-1	0	1	2	3	4	≥ 5
≤ -5	0	0	0	0	1	0	0	0	0	0	0
-4	0	2	2	2	0	2	1	1	0	0	0
-3	0	2	2	2	2	2	0	1	0	0	0
-2	4	1	4	15	24	18	14	3	4	1	0
-1	7	14	47	80	168	240	147	76	25	9	1
0	25	37	103	260	455	688	595	325	153	65	23
1	6	24	68	162	317	538	501	331	180	65	40
2	0	1	13	27	59	98	119	112	82	39	27
3	0	0	0	1	5	12	24	18	21	14	13
4	0	0	0	0	0	0	3	0	2	0	0
≥ 5	0	0	0	0	0	0	0	0	0	0	0

BOO

Prediction \ Observation	≤ -5	-4	-3	-2	-1	0	1	2	3	4	≥ 5
≤ -5	0	0	0	0	1	0	0	0	0	0	0
-4	0	2	2	2	0	2	1	1	0	0	0
-3	0	1	2	8	9	6	3	2	1	0	0
-2	4	3	14	27	45	60	43	13	5	3	0
-1	13	15	50	94	189	267	194	108	40	17	3
0	17	30	90	217	410	627	501	276	151	57	23
1	8	26	63	164	290	489	475	306	158	55	35
2	0	2	14	34	75	129	148	126	75	39	28
3	0	0	1	1	10	12	31	31	28	16	13
4	0	0	1	0	1	4	7	3	6	5	2
≥ 5	0	0	0	0	0	0	0	0	3	1	0

SVM

Prediction \ Observation	≤ -5	-4	-3	-2	-1	0	1	2	3	4	≥ 5
≤ -5	0	0	0	0	0	0	0	0	0	0	0
-4	0	1	1	2	1	3	0	1	0	0	0
-3	0	2	5	6	11	8	6	2	0	0	0
-2	4	5	16	38	44	68	40	17	7	2	1
-1	11	20	55	107	218	280	220	113	54	20	3
0	22	28	84	192	379	593	496	281	131	48	25
1	5	18	60	159	280	463	419	272	155	61	32
2	0	5	12	37	81	144	170	135	80	36	25
3	0	0	4	6	16	32	42	39	32	21	16
4	0	0	0	0	0	5	10	6	7	5	2
≥ 5	0	0	0	0	0	0	0	0	1	0	0

LIR

Prediction \ Observation	≤ -5	-4	-3	-2	-1	0	1	2	3	4	≥ 5
≤ -5	0	0	1	0	0	1	1	0	0	0	0
-4	0	0	1	2	0	1	1	1	0	0	0
-3	3	2	4	14	19	11	7	3	2	0	0
-2	6	7	21	33	52	69	35	16	6	4	1
-1	11	21	54	102	202	240	173	99	44	12	4
0	15	30	81	189	349	570	474	267	121	54	20
1	6	18	61	158	297	527	499	313	172	69	46
2	1	1	13	46	97	161	178	148	95	41	25
3	0	0	1	3	13	16	31	18	23	12	8
4	0	0	0	0	1	0	4	1	4	1	0
≥ 5	0	0	0	0	0	0	0	0	0	0	0

strategy. Accuracy is defined as the percentage of correctly predicted outcomes—possible outcomes are "home team wins", "draw", and "away team wins". Root mean squared error (RMSE) represents the square root of the second sample moment of the differences between predicted values and observed values. Mean absolute deviation (MAD) is the average of the absolute deviations of predicted values and observed values. Both the RMSE and MAD are calculated based on the (predicted) goal difference of home team and away team. In terms of prediction quality, RFO achieves the best results, i.e. the highest accuracy, the lowest RMSE, and the lowest MAD. RFO is followed by the other two machine learning methods BOO and SVM. After this, we find the less complex LIR followed by the unsophisticated approaches BET and HOM. It is not surprising that the random bet RAN clearly leads to the poorest performance.

Another finding is that the accuracy of the different approaches is in line with the resulting average payoff. In addition, the machine learning approaches (RFO, BOO, SVM) reveal average payoffs greater than 1, in contrast to the other less complex methods. In this context it should be noted that an average payoff greater than 1 means that the betting party beats the bookmaker on a long term perspective. Like the prediction quality, the average payoffs follow the complexity of the corresponding methods.

Moreover, betting that the home team wins is clearly preferred by all approaches except RAN. This makes sense since RAN is supposed to bet on the home team in about half of the cases. All other methods identified the home advantage we described in Sect. 3. This is somehow reasonable as the home advantage can be taken into account by the statistical models. Since the bookmakers' odds, which are on average lower for the home teams than for the away teams, are not part of the data set, the strategies mostly vote for the home team. Furthermore, we find that a higher number of bets leads to lower average payoffs. It could be carefully concluded that the machine learning methods are better at selecting the more secure match outcomes out of the pool

Table 5 Risk-return characteristics per match for the strategies RFO, BOO, SVM, LIR, BET, HOM, and RAN from season 2013/14 to 2017/18

	RFO	BOO	SVM	LIR	BET	HOM	RAN
Mean	0.0542	0.0217	0.0054	− 0.0043	− 0.0295	− 0.0324	− 0.0630
p-value of WT	0.0028	0.1108	0.8439	0.3235	0.0000	0.0000	0.0000
Minimum	− 1.0000	− 1.0000	− 1.0000	− 1.0000	− 1.0000	− 1.0000	− 1.0000
Quartile 1	0.0375	− 1.0000	− 1.0000	− 1.0000	− 1.0000	− 1.0000	− 1.0000
Median	0.1650	0.1800	0.1700	0.1800	0.2000	− 1.0000	− 1.0000
Quartile 3	0.3525	0.4000	0.4400	0.5000	0.8000	0.7800	0.7200
Maximum	5.2500	6.0000	6.0000	5.2500	1.8500	13.5000	20.0000
Standard deviation	0.7545	0.7816	0.8844	0.8818	0.9433	1.2993	1.5510
Skewness	1.2106	0.9184	1.2001	0.7542	0.1608	2.1598	3.0134
Kurtosis	7.8638	6.4619	5.3599	2.3316	− 1.5890	9.9730	18.4355
Share with return > 0	0.7562	0.7095	0.6612	0.6360	0.5449	0.4610	0.3782

WT denotes the non-parametric Wilcoxon–Test

of all possible bets. Finally, we notice that the minimum and maximum of the predicted goal difference between home team and away team goals are in a reasonable range. The machine learning approaches in particular lead to values between 4.9 goals difference for the home team and 4.6 goals difference for the away team.

In Fig. 4, the results of the simulation study are presented in form of a contingency table comparing the observed and the predicted goal difference for each considered machine learning algorithm. A perfect prediction would produce a matrix where only the diagonal elements form the upper left to the lower right corner are greater than 0. Values on the lower left or the upper right would express a bad prediction quality. In general we can find, the most of the predictions (as well as the observed outcomes) concentrate around a draw (goal difference = 0). The algorithms do not tend to produce extreme predictions, like goal differences of more than three goals. In addition, the elements of the lower left or the upper right equal 0. For the algorithms BOO, SVM, and LIR, the home advantage is incorporated in the predictions since the matrices are more dense in the lower half. RFO is even able to identify some matches the away team clearly wins.

5.2 Financial Analysis

Table 5 depicts the risk-return characteristics per match of the seven considered strategies. First of all, we observe that the findings from the statistical analysis (see Sect. 5.1) are also reflected in the finance context—strategies with a high prediction quality generate meaningful profits and vice versa. The approaches based on machine learning algorithms provide positive returns ranging between 0.5% per match for SVM and 5.42% per match for RFO. In contrast, the benchmark methods achieve negative returns between −0.43% (LIR) and −3.24% (HOM). As expected, the naive strategy RAN produces a clear loss of −6.30% per match. Applying RFO results in economically and statistically significant returns—the *p* value of the non-parametric Wilcoxon–Test (WT) is 0.28%. It is not surprising that the minimum of all strategies is −1, as at least one predicted event does not occur. Analyzing quartile 1 confirms the outperformance of RFO—more than 75% of all returns are in the positive range. The maximum is vastly different between RFO, BOO, SVM, LIR, and BET on the one hand and HOM and RAN on the other hand. RFO achieves the highest hit ratio, i.e. the percentage of matches with positive returns, with 75.62%. In summary, the strategies based on machine learning methods outperform classic approaches in a multitude of return characteristics and risk metrics—this statement is particularly true for RFO.

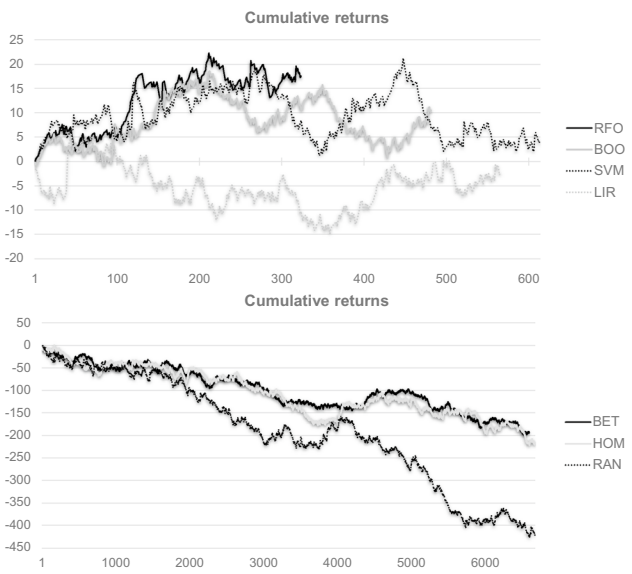

Fig. 5 Development of cumulative returns of the strategies RFO, BOO, SVM, and LIR (upper graph) as well as BET, HOM, and RAN (lower graph) from season 2013/14 to 2017/18

Following Knoll et al. [31] and Stübinger [54], we analyze the performance of the strategies over time. Figure 5 shows the cumulative returns of RFO, BOO, SVM, LIR (upper graph) and BET, HOM, RAN (lower graph) from 2013/14 to 2017/18. The difference in the length of the seven time series is caused by the varying number of executed bets (see Table 4). As expected, RFO is the best in class with an end value of 17.55—the smooth and steady growth is particularly pleasant for any potential investor. BOO displays medium drawdowns and a final cumulative return of 10.46. For SVM and LIR, gains and losses offset each other over time. In contrast, the remaining strategies possess a steady race to the bottom. As a result, BET and HOM achieve a cumulative return of −194.60 and −216.11 at the end of season 2017/18. The naive strategy RAN performs even worse: the loss totals −420.11 within the considered time period. In summary, the findings obtained so far are not driven by outliers, but rather the result of permanent correct or incorrect predictions.

Figure 6 examines the risk aversion of the individual strategies by showing the relative proportions of the executed bets. For each strategy, the betting odds are divided into the following five classes: low (1.00–1.50), low-medium (1.51–2.00), medium (2.01–3.00), medium-high (3.01–4.00), high (> 4.01). First of all, a clear asymmetry between strategies based on dependent variables and strategies based on betting odds can be identified. RFO selects low betting odds in approximately 69% of all executed bets—low-mid betting odds are used in 20%. One could conclude that this approach relies to a high degree on "safe" outputs and avoids risky bets. The strategies BOO, SVM, and LIR also tend to focus more on lower odds, but the willingness to take risks steadily increases. BET only applies low, low-medium, and medium betting odds because one of the odds for "home team wins" or "away team wins" is always less than 3.00. As expected, RAN chooses the trading rule independently of the available betting odds – matches with low or high betting odds are chosen similarly often. In conjunction with Table 5, it

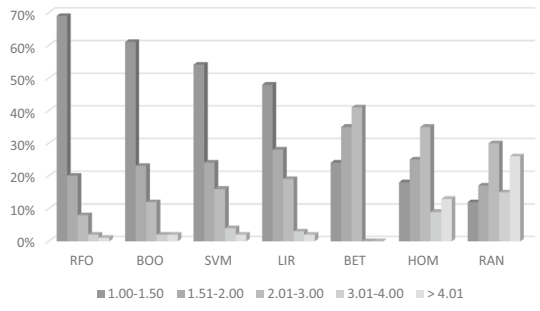

Fig. 6 Relative proportions of used betting odds of the strategies RFO, BOO, SVM, and LIR as well as BET, HOM, and RAN from season 2013/14 to 2017/18

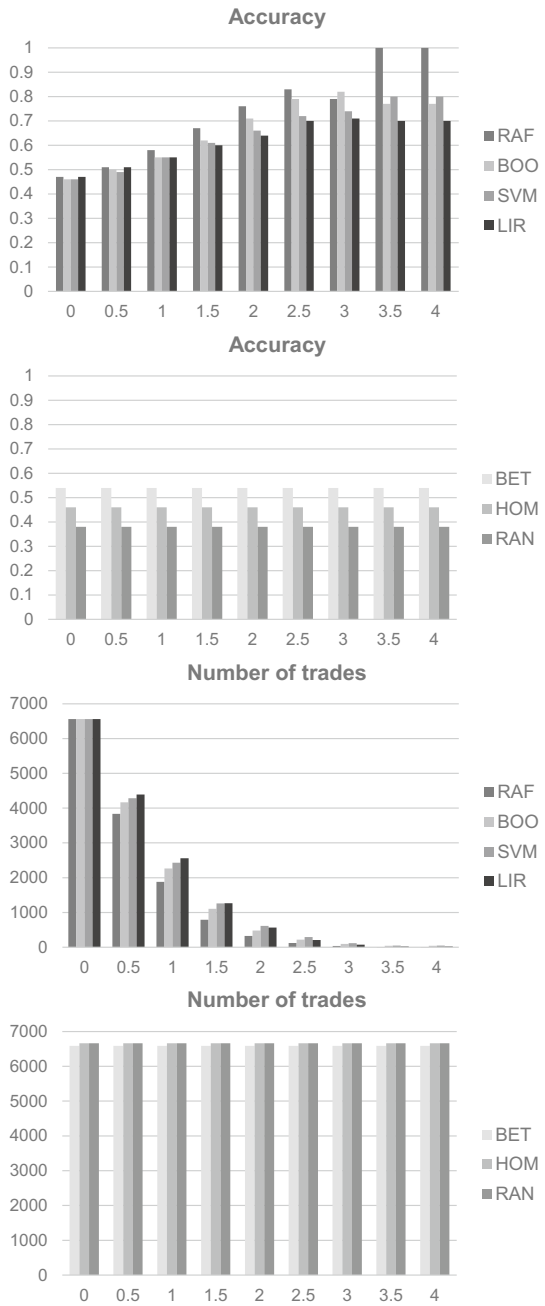

Fig. 7 Accuracy and number of bets in relation to the trading threshold for each approach

can be summarized that risk-averse strategies lead to the most profitable returns—this finding is well in line with the literature ([15, 16, 57]).

Figure 7 presents the accuracy and the number of bets conducted in relation to the trading threshold. Remember, the trading thresholds take on the task of selecting the matches where a larger goal difference was predicted (assuming that the outcome of these matches could be predicted more precisely). For RAF, BOO, SVM and LIR,

the accuracy of the predicted outcome (home win, draw, away win) increases with a higher trading threshold which indicates that a larger predicted goal difference results in a more reliable prediction of the outcome. On the other hand, the number of bets conducted decreases for higher trading thresholds which means less bets are conducted in our simulation and therefore the possible total profit decreases. Consequently in terms of maximizing the profit, there has to be a trade-off between the accuracy of the predicted outcome generated by a higher trading threshold and the potential total profit caused by a rather low trading threshold. For the sake of completeness, we draw the same diagrams for the approaches BET, HOM, and RAN. The values stay the same with a varying trading threshold because these approaches do not predict a goal difference and thus are not capable of including a trading threshold.

6 Conclusions and Future Work

In this paper, we have presented a data-driven approach for predicting the outcome of football league matches and generating positive returns by betting accordingly. These positive returns could be reached by applying machine learning algorithms to large data sets. The back-testing study revealed that the random forest approach generated statistically and economically significant returns of 5.42% on average for each bet. In contrast, less complex approaches, namely linear regression, placing all bets on the home team, betting on the (most likely) outcome with the lowest betting odd, or choosing a random match result, were not able to generate positive returns. Moreover, it turned out that risk-averse strategies yield higher returns and that a higher trading threshold (selecting matches in which a larger goal difference was predicted) leads to a more reliable prediction of the outcome.

This study could serve as a starting point for future work. It could be used as a reference for research applying data-driven approaches to forecast the result of sporting events. For example, other types of sports with corresponding leagues, such as rugby, American football, or basketball could be targeted similarly.

Acknowledgements We are grateful to two anonymous referees for many helpful suggestions on this topic.

References

1. Archontakis F, Osborne E (2007) Playing it safe? A Fibonacci strategy for soccer betting. J Sports Econ 8(3):295–308
2. Avellaneda M, Lee JH (2010) Statistical arbitrage in the US equities market. Quant Finance 10(7):761–782
3. Bernile G, Lyandres E (2011) Understanding investor sentiment: the case of soccer. Financ Manag 40(2):357–380
4. Bertram WK (2010) Analytic solutions for optimal statistical arbitrage trading. Phys A Stat Mech Appl 389(11):2234–2243
5. Bollinger J (2001) Bollinger on bollinger bands. McGraw-Hill, New York
6. Boshnakov G, Kharrat T, McHale IG (2017) A bivariate weibull count model for forecasting association football scores. Int J Forecast 33(2):458–466
7. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
8. Bühlmann P, Hothorn T (2007) Boosting algorithms: regularization, prediction and model fitting. Stat Sci 22(4):477–505
9. Chen T, He T, Benesty M (2015) xgboost: extreme gradient boosting. R package version 0.3-0. In: TechnicalReport
10. Choi D, Hui SK (2014) The role of surprise: understanding overreaction and underreaction to unanticipated events using in-play soccer betting market. J Econ Behav Org 107:614–629
11. Croxson K, Reade J (2014) Information and efficiency: goal arrival in soccer betting. Econ J 124(575):62–91
12. Dixon M, Coles S (1997) Modelling association football scores and inefficiencies in the football betting market. J R Stat Soc Ser C (Appl Stat) 46(2):265–280
13. Dragulescu AA, Dragulescu MAAA (2014) PROVIDE, R. Package 'xlsx'. Cell, 2018, 9. Jg., Nr. 1, S. 5
14. Egidi L, Pauli F, Torelli N (2018) Combining historical data and bookmakers' odds in modelling football scores. Stat Model 18(5–6):436–459
15. Endres S, Stübinger J (2019) Optimal trading strategies for Lévy-driven Ornstein–Uhlenbeck processes. Appl Econ 51(29):3153–3169
16. Endres S, Stübinger J (2019) Regime-switching modeling of high-frequency stock returns with Lévy jumps. Quantitative Finance, Forthcoming
17. Forrest D, Simmons R (2008) Sentiment in the betting market on Spanish football. Appl Econ 40(1):119–126
18. Franck E, Verbeek E, Nüesch S (2010) Prediction accuracy of different market structures—bookmakers versus a betting exchange. Int J Forecast 26(3):448–459
19. Franck E, Verbeek E, Nüesch S (2013) Inter-market arbitrage in betting. Economica 80(318):300–325
20. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat: 1189–1232
21. Gatev E, Goetzmann WN, Rouwenhorst KG (2006) Pairs trading: performance of a relative-value arbitrage rule. Rev Financ Stud 19(3):797–827
22. Gil RGR, Levitt SD (2012) Testing the efficiency of markets in the 2002 World Cup. J Predict Markets 1(3):255–270
23. 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. In: KDD workshop on large-scale sports analytics, New York, USA, 24–28 Aug 2014
24. Groll A, Kneib T, Mayr A, Schauberger G (2018) On the dependency of soccer scores—a sparse bivariate poisson model for the UEFA European football championship 2016. J Quant Anal Sports 14(2):65–79
25. Groll A, Ley C, Schauberger G, Van Eetvelde H (2019) A hybrid random forest to predict soccer matches in international tournaments. J Quant Anal ports. (to appear)
26. Groll A, Schauberger G, Tutz G (2015) Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: an application to the FIFA World Cup 2014. J Quant Anal Sports 11(2):97–115
27. Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. Math Intelligencer 27(2):83–85

28. Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15(3):651–674
29. Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: implications for stock market efficiency. J Finance 48(1):65–91
30. Kelly AH (1956) The fourteenth amendment reconsidered: the segregation question. Mich Law Rev 54(8):1049–1086
31. Knoll J, Stübinger J, Grottke M (2019) Exploiting social media with higher-order factorization machines: statistical arbitrage on high-frequency data of the S&P 500. Quant Finance 19(4):571–585
32. Koopman EME, Hakemulder F (2015) Effects of literature on empathy and self-reflection: a theoretical-empirical framework. J Lit Theory 9(1):79–111
33. Leifeld P (2013) texreg: conversion of statistical model output in R to HTML tables. J Stat Softw 55(8):1–24
34. Levitt SD (2004) Why are gambling markets organised so differently from financial markets? Econ J 114(495):223–246
35. Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22
36. Lisi F, Zanella G (2017) Tennis betting: can statistics beat bookmakers? Electron J Appl Stat Anal 10(3):790–808
37. Liu B, Chang LB, Geman H (2017) Intraday pairs trading strategies on high frequency data: the case of oil companies. Quant Finance 17(1):87–100
38. Luckner S, Schröder J, Slamka C (2008) On the forecast accuracy of sports prediction markets. Negotiation, auctions, and market engineering. Springer, Berlin, Heidelberg, pp 227–234
39. Maher M (1982) Modelling association football scores. Stat Neerl 36(3):109–118
40. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2017) e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package version 1.6-8
41. Palomino F, Renneboog L, Zhang C (2009) Information salience, investor sentiment, and stock returns: the case of British soccer betting. J Corp Finance 15(3):368–387
42. Peterson BG, Carl P, Boudt K, Bennett R, Ulrich J, Zivot E, Wuertz D (2014) Performance analytics: econometric tools for performance and risk analysis. R package version 1.4. 3541
43. Pfaff B, McNeil A, Ulmann S (2013) QRM: provides R language code to examine quantitative risk management concepts. R package version 0.4-9. http://CRAN.R-project.org/package=QRM
44. R Core Team (2017) stats: a language and environment for statistical computing. R package
45. Team RC, Wuertz D, Setz T, Chalabi Y (2015) timeSeries: Rmetrics —Financial time series objects. R package version, 3012
46. Rue H, Salvesen O (2000) Prediction and retrospective analysis of soccer matches in a league. J R Stat Soc Ser D (Stati) 49(3):399–418
47. Ryan JA, Ulrich JM (2017) quantmod: Quantitative financial modelling framework. R package version 0.4-12
48. Ryan JA, Ulrich JM (2014) xts: eXtensible time series. R package version 0.8-2
49. Schauberger G, Groll A, Tutz G (2018) Analysis of the importance of on-field covariates in the German Bundesliga. J Appl Stat 45(9):1561–1578
50. Spann M, Skiera B (2009) Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. J Forecast 28(1):55–72
51. Stefani RT (1980) Improved least squares football, basketball, and soccer predictions. IEEE Trans Syst Man Cybernetics 10(2):116–123
52. Steinwart I, Christmann A (2008) Support vector machines. Springer, New York
53. Stekler HO, Sendor D, Verlander R (2010) Issues in sports forecasting. Int J Forecast 26(3):606–621
54. Stübinger J (2019) Statistical arbitrage with optimal causal paths on high-frequency data of the S&P 500. Quant Finance 19(6):921–935
55. Stübinger J, Endres S (2018) Pairs trading with a mean-reverting jump-diffusion model on high-frequency data. Quant Finance 18(10):1735–1751
56. Stübinger J, Knoll J (2018) Beat the bookmaker - Winning football bets with machine learning (Best Application Paper). In: proceedings of the 38th SGAI international conference on artificial intelligence, pp. 219–233. Springer
57. Stübinger J, Mangold B, Krauss C (2018) Statistical arbitrage with vine copulas. Quanti Finance 18(11):1831–1849
58. Tax N, Joustra Y (2015) Predicting the Dutch football competition using public data: a machine learning approach. Trans Knowl Data Eng 10(10):1–13
59. Trapletti A, Hornik K, Lebaron B (2007) Tseries: time series analysis and computational finance. R package version 0.10-11
60. Ulrich J (2016) TTR: technical trading rules. R package
61. Wickham H, Bryan J (2016) readxl: Read Excel files. R package 1.0. 0. 2017
62. Wickham H, Francois R, Henry L, Müller K (2015) dplyr: a grammar of data manipulation. R package version 0.4, 3
63. Wickham H, Hester J, Francois R, Jylänki J, Jørgensen M (2017) readr: read rectangular text data. R foundation for statistical computing. R package version 1.1.1
64. Zeileis A (2006) Object-oriented computation of sandwich estimators. J Stat Softw 16(9):1–16
65. Zeileis A, Grothendieck G (2005) zoo: S3 infrastructure for regular and irregular time series. J Stat Softw 14(6):1–27
66. Zeileis A, Leitner C, Hornik K (2016) Predictive bookmaker consensus model for the UEFA Euro 2016. In: Working papers in economics and statistics
67. Zeileis A, Leitner C, Hornik K (2018) Probabilistic forecasts for the 2018 FIFA World Cup based on the bookmaker consensus model. In: working papers in economics and statistics
68. Zhou ZH (2012) Ensemble methods: foundations and algorithms. Chapman and Hall, Boca Raton