International Journal of Forecasting 38 (2022) 895–909

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting football results and exploiting betting markets: The case of "both teams to score"

Igor Barbosa da Costa [a,*], Leandro Balby Marinho [b], Carlos Eduardo Santos Pires [b]

[a] Federal Institute of Education, Science, and Technology of Paraiba (IFPB), Brazil
[b] Federal University of Campina Grande (UFCG), Brazil

ARTICLE INFO

Keywords:
Football
Soccer prediction
Sports betting
Machine learning
Forecasting
Feature engineering

ABSTRACT

The continuous growth of available football data presents unprecedented research opportunities for a better understanding of football dynamics. While many research works focus on predicting which team will win a match, other interesting questions, such as whether both teams will score in a game, are still unexplored and have gained momentum with the rise of betting markets. With this in mind, we investigate the following research questions in this paper: "How difficult is the 'both teams to score' (BTTS) prediction problem?", "Are machine learning classifiers capable of predicting BTTS better than bookmakers?", and "Are machine learning classifiers useful for devising profitable betting strategies in the BTTS market?". We collected historical football data, extracted groups of features to represent the teams' strengths, and fed these to state-of-the-art classification models. We performed a comprehensive set of experiments and showed that, although hard to predict, in some scenarios it is possible to outperform bookmakers, which are robust baselines per se. More importantly, in some cases it is possible to beat the market and devise profitable strategies based on machine learning algorithms. The results are encouraging and, besides shedding light on the problem, may provide novel insights for all kinds of football stakeholders.

1. Introduction

Association football, commonly known as football or soccer, is a sport played by more than 250 million people in over 200 countries, making it the world's most popular sport (FIFA, 2017). Trying to predict the performance of teams or the occurrence of events during a match is commonplace among football enthusiasts and professional circles. In recent decades, researchers have proposed different methods for understanding predictability and randomness in football match outcomes (Anderson & Sally, 2013). These results are relevant to all people involved in the sport. For instance, managers, coaches, and players can analyze team performance in a more informed way; punters can make better decisions on betting markets; and sports press and fans can better understand the characteristics and patterns of the game.

Sports forecasting refers to drawing conclusions about future performance based on the combined interactions of previously gathered information, knowledge, or data (Hughes & Franks, 2015). Football forecasting is an especially difficult task, given the strong stochastic nature of this kind of sport. Besides the skills of the players, there are countless features, such as team morale, injuries, current score, and rank, among others, which may influence the course of a match.

Nowadays, there are several publicly available datasets on the Web providing information about matches, players, teams, and championships, thus presenting unprecedented

* Corresponding author.
E-mail addresses: igor.costa@ifpb.edu.br (I.B. da Costa), lbmarinho@dsc.ufcg.edu.br (L.B. Marinho), cesp@dsc.ufcg.edu.br (C.E.S. Pires).

https://doi.org/10.1016/j.ijforecast.2021.06.008
0169-2070/

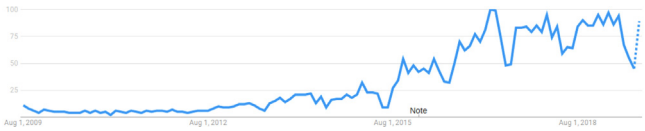

Fig. 1. Interest in BTTS over time (2009–2019) according to Google Trends.

research opportunities for better understanding football dynamics. Data science and machine learning (ML) have often been employed to handle the volume and complexity of such datasets (Berrar, Lopes, Davis, & Dubitzky, 2018a).

The vast majority of existing research on football forecasting focuses on predicting which team will win a particular match or tournament. Other interesting problems—such as "How many goals will be scored in a match?", "Which team will have more ball possession?", and "When will the first goal be scored?"—still receive little attention from the research community. Besides raising debate among fans, questions like these have gained significant importance with the recent boom of online sports betting markets around the world.

The global sports betting market is currently estimated to be worth up to $3 trillion, where football betting represents 65% of this value (Dailymail, 2015). Bookmakers offer odds (i.e., the ratio of payoff to stake) for punters regarding, besides the questions mentioned above, whether both teams will score in a particular match. This question—called "Both teams to score?" (BTTS)—is one of the most popular in football betting (Betsonly, 2017) and has aroused growing interest in the last decade (see Fig. 1).

In this paper, we cast BTTS as a binary classification problem where we want to predict whether both teams will score or not. By studying this problem under a data science/machine learning view, we expect to shed light on the BTTS problem, not only from an ML/sports forecasting perspective but also from a sports betting market perspective by evaluating the efficiency of betting strategies in the BTTS market. There is a wide range of work in economic literature on football betting market efficiency. In short, if the market is efficient, we should not be able to devise profitable systematic strategies. Thus, we assess this assumption from an ML perspective, where we evaluate whether carefully engineered features fed to state-of-art classifiers can be useful for devising such strategies in the BTTS market. In this way, more specifically, we intend to answer the following research questions:

- RQ1: How difficult is the BTTS prediction problem?
- RQ2: Are ML classifiers capable of predicting BTTS better than bookmakers?
- RQ3: Are ML classifiers useful for devising profitable betting strategies in the BTTS market?

To this end, we built Web crawlers to collect football match results from six seasons (2013–2019) of nine of the most prestigious championships in the world (see Table 1). These crawlers collected data on published odds from 19 bookmakers for these same matches. Given all this collected data, we extracted several features and fed them to several state-of-the-art ML classifiers. Finally, we conducted a thorough set of experiments where we compared the outcomes of the classifiers with the forecasts made by bookmakers.

Table 1
List of championships in the dataset.

Country	Championship	Label
Brazil	Serie A	Brazil A
Brazil	Serie B	Brazil B
England	Premier League	England A
France	Ligue 1	France A
Germany	Bundesliga	Germany A
Italy	Serie A	Italy A
Netherlands	Eredivisie	Netherlands A
Spain	La Liga	Spain A
Portugal	Primeira Liga	Portugal A

Among the main findings, we discovered that (i) distinct championships have different levels of predictability, (ii) the class distribution is different across the championships, (iii) some classifiers can outperform bookmakers' predictions, and (iv) some classifiers can devise profitable betting strategies.

The remainder of the paper is structured as follows. In Section 2, we present important concepts related to the sports betting market. In Section 3, we present related works and highlight the differences and contributions of our research in comparison to those. In Section 4, we describe the collected data and detail the feature engineering process we carried out on these data. In Section 5, we detail all the classifiers evaluated in this work. In Section 6, we conduct a thorough set of experiments and analysis for answering the research questions that drove our research. Finally, in Section 7, we summarize the main contributions of this work, emphasizing opportunities for future work.

2. Football betting market

To provide a better understanding of this work, in this section we review the concepts behind the "odds" in football betting markets.

Bookmakers post betting odds, which can be defined as the ratio of payoff for a stake. The odds value is always greater than 1, since this value represents the bet value itself. For instance, if we bet on a result with odds of 2.5 and our bet succeeds, we will get back 2.5 times the stake value. The smaller the odds, the higher the probability of the event happening according to the bookie (favorite) and conversely for higher odds (underdog).

There are several types of markets available for betting. In our work, we collected odds from the Money Line market and Both Teams to Score (BTTS) market. The Money Line market offers odds for three possible results: home

team win (H), away team win (A), and draw (D), while the BTTS market offers odds on Yes (Y), both teams scoring a goal, or No (N), at least one of them failing to score.

The odds values, to a certain extent, represent the probabilities predicted by the bookmakers for each possible outcome, as well as a profitable safety margin for them. For instance, consider a football match with BTTS odds of 2.25 for Yes, and 1.61 for No. Inverse odds are an indication of the bookmaker's degree of belief. In this case, $\frac{1}{2.25} = 0.44$, i.e., the chance of both teams to score is 44%. However, since bookmakers do not offer perfectly fair odds, the sum of these numbers (also known as the booksum, bookmaker take, or bookmaker margin) will always be greater than 1 ($\frac{1}{2.25} + \frac{1}{1.61} = 1.06$). This margin exceeds 1 to ensure that the bookmaker has a positive expected value (profit).

Given that the sum of inverse odds is often higher than 1, it is necessary to normalize these values to extract the real probabilities. Many studies use standard normalization to handle the booksum, dividing the inverse odds by the booksum. However, (Štrumbelj, 2014) demonstrated that Shin's method (Shin, 1993) is more accurate than standard normalization or regression-based models. Thus, in this paper, we adopt Shin's method to convert odds into probabilities.

3. Related works

Investigating inherent patterns in football data is not a new area of research. Back in the mid-1960s, Moroney (1962) proposed the use of a Poisson distribution for fitting the number of goals per match. Since then, researchers have made significant advances towards improving forecasts related to football match outcomes. In general, these approaches follow two different strategies. In the first one, the models predict the number of goals scored by each team, to then derive the probabilities of each result (home win, draw, or away win) (Angelini & De Angelis, 2017; Dixon & Coles, 1997; Karlis & Ntzoufras, 2003; Rue & Salvesen, 2000). In the second strategy, the models estimate these probabilities directly, without necessarily estimating the number of goals of each team (Forrest, Goddard, & Simmons, 2005; Goddard, 2005; Hubáček, Šourek, & Železnỳ, 2018). An interesting overview of different forecasting methods is given by Spann and Skiera (2009). To the best of our knowledge, there is no previous study focused strictly on the BTTS problem, but one could use models that predict the number of goals scored (the first strategy) to derive predictions of BTTS.

Most of these models assume that the pairwise observation (scoreboard) comes from a bivariate distribution. Hence, they express the probability of any possible outcome as $P(X = x \; and \; Y = y)$, where X and Y denote the number of goals scored by the home and away teams, respectively. For the BTTS problem in particular, we adapt these models to forecast Yes or No, i.e., $P(X > 0 \; and \; Y > 0)$ and $P(X = 0 \; or \; Y = 0)$, respectively.

Regarding the modeling, the parameters of the bivariate distribution can be expressed as a function of the team's inherent attacking and defensive strengths. This approach was originally proposed by Maher (1982), who adopted a double-Poisson distribution as the underlying distribution for goal scoring. Dixon and Coles (1997) expanded that model for low-scoring games, introducing a dependence parameter for the match results 0–0, 1–0, 0–1, and 1–1. They also proposed a weighting function to give more importance to the most recent matches. Rue and Salvesen (2000) adapted the framework of Dixon and Coles (1997) and developed a dynamic generalized linear model analyzed by Markov Chain Monte Carlo in continuous time. Karlis and Ntzoufras (2003) also used a bivariate Poisson distribution and showed that the introduction of a parameter for dependence between the goals scored by a team leads to a more accurate prediction of the number of draws. Goddard (2005) compared the efficiency of a bivariate Poisson distribution for predicting goals against an ordered probit regression for estimating the results directly. Koopman and Lit (2015) used a state-space model to allow team strengths to change stochastically over time. Angelini and De Angelis (2017) proposed an innovative approach based on Poisson autoregression with exogenous covariates (PARX). Boshnakov, Kharrat, and McHale (2017) presented a Weibull inter-arrival-times-based count process and a copula to produce a bivariate distribution of the numbers of goals scored.

Our proposal is in line with the second strategy. That is, we use ML classifiers to predict the probabilities of each result directly. There is a wide variety of ML techniques employed in the literature, such as Bayesian networks (Constantinou, 2019; Joseph, Fenton, & Neil, 2006; Min, Kim, Choe, Eom, & McKay, 2008), support vector machines (Igiri, 2015; Tüfekci, 2016), logistic regression (Hubáček, Šourek, & Železnỳ, 2019; Prasetio et al., 2016), neural networks (Bunker & Thabtah, 2017; Hubáček et al., 2019), k-nearest neighbors (Berrar, Lopes, & Dubitzky, 2018b; Hucaljuk & Rakipović, 2011), and gradient boosting, which have been used to offer some of the best results for football match prediction (Berrar et al., 2018b; Hubáček et al., 2018). An interesting overview of the performances of ML classifiers in modeling football results is given by Baboota and Kaur (2019).

In 2017, the journal *Machine Learning* (Springer) organized a special issue and challenge on football analytics. The task was to predict the outcomes of future matches based on a dataset of over 200,000 football matches from football leagues around the world (Berrar, Lopes, Davis, & Dubitzky, 2017). The conclusion from the best approaches of this challenge pointed out that a critical factor for getting good predictions is the ability to leverage domain knowledge (Berrar et al., 2018b). In other words, a good feature engineering process is key to good predictions in the football domain.

The best models of this challenge present a great variety of ideas for encoding domain knowledge in the prediction models (Berrar et al., 2018b). Among these ideas, we can highlight team ratings, which are models developed to represent the strength of teams according to one or a few variables. Hvattum and Arntzen (2010), Leitner, Zeileis, and Hornik (2010) used a variation of the Elo model, which was initially created to evaluate chess players. Constantinou (2019) used the pi-rating, which

provides relative measures of superiority between football teams based on the relative discrepancies in scores between opponents. Hubáček et al. (2018) used not only the pi-rating but also a rating based on an adaptation of the well-known PageRank algorithm.

Although most of the literature focuses on the development of more accurate prediction models, there is growing interest in using these models to devise profitable betting strategies (Constantinou, 2019). Studies in this direction are essential, since they put the efficient-market hypothesis to the test. Several studies have shown that it is possible to obtain consistent profits in the betting market from strategies that combine the prediction of statistical models with econometric approaches. Hubáček et al. (2019) designed a strategy for bet distribution according to the odds and model predictions, trading off profit expectation and variance. Koopman and Lit (2019) used a positive "expected value" strategy, in which one bet is wagered when the probability predicted by the model is higher than the implied probabilities of odds. Boshnakov et al. (2017) used the famous Kelly criterion (Kelly, 2011) that, in addition to verifying whether a bet has a positive expected value, defines the amount to be wagered based on the difference in predictions between the model and the market. An overview of the efficiency of online football betting markets is given by Angelini and De Angelis (2019).

In summary, previous works present promising strategies for tackling the problem of predicting football outcomes and its implications for the betting market. We bring some novel contributions to the field. As with Boshnakov et al. (2017), Gomes, Portela, and Santos (2016) and Owen (2017), our aim is not to predict the winner of a match. We tackle the BTTS problem, which to our knowledge has not yet been considered in the literature. We perform careful feature engineering and evaluate several classifiers exploiting these features, similar to how most researchers on football match outcome prediction tackle the problem (Berrar et al., 2018b; Tüfekci, 2016). We also evaluate the combination of the use of bookmakers' odds with historical data to make predictions, similar to Egidi, Pauli, and Torelli (2018) and Titman, Costain, Ridall, and Gregory (2015). Finally, we assess the classifiers in terms of both accuracy and profitability, considering the BTTS betting market. We do that by evaluating popular betting strategies, such as the "expected value" (Koopman & Lit, 2019) and adaptations of the Kelly criterion (Boshnakov et al., 2017).

4. Data preprocessing

This section describes in detail the collected data, as well as the features we extracted from it. The classification algorithms use these features for the BTTS prediction problem.

4.1. Experimental datasets

During our investigation, we created Web crawlers to collect data from betexplorer.com, one of the most popular sites in the world for sports betting. These crawlers indexed the match outcomes and the odds of several bookmakers. After collecting these data, we scraped them (HTML format) and generated two structured datasets.

In the first dataset, we stored the match results of six seasons (2013–2019) from nine championships (see Table 1). In this paper, this dataset is referred to as the Match dataset. In the second dataset, for the same matches as the Match dataset, we stored the odds values (the last value before the matches start) from 19 Web bookmakers regarding two markets: Money Line (ML) and Both Teams to Score (BTTS). In this study, this dataset is referred to as the Odds dataset. Table 2 displays the data fields of both datasets, and Table 3 summarizes some basic statistics about them.

Brazilian and European leagues have different calendars. In Brazil, a season starts and ends in the same year, while in Europe the season begins in one year and finishes in the next. Thus, to standardize across the different leagues, we identified the European championships by the year the season started. For example, the 2016 season refers to the 2016–2017 season.

Table 2
Data fields collected.

Match dataset	Odds dataset
Championship ID	Championship ID
Season ID	Bookmaker ID
Match ID	Match ID
Round Number	Odds for Home Team (ML)
Match Date	Odds for Away Team (ML)
Home Team ID	Odds for Draw (ML)
Away Team ID	Odds for Yes (BTTS)
Home Team Goals	Odds for No (BTTS)
Away Team Goals	

Table 3
Dataset: summary statistics.

Championship	# Matches	# Odds BTTS	# Odds ML
Brazil A	2280	32,043	46,769
Brazil B	2280	29,215	46,587
Portugal A	1770	26,154	36,500
France A	2280	34,406	47,754
England A	2280	36,130	47,938
Spain A	2280	34,618	47,806
Italy A	2280	34,556	47,832
Germany A	1836	27,914	38,484
Netherlands A	1836	26,647	37,654
Total	19,122	281,683	397,324

Odds BTTS: Number of BTTS odds; # Odds ML: Number of Money Line odds.

4.2. Feature engineering

Here, we describe in detail how we modeled the features to be used by the chosen classifiers. We extracted specific features from the Match and Odds datasets. In the former, the features are more related to team performance while in the latter they are about the market's "opinion" through their odds.

4.2.1. Target variable

We start by defining the target variable. This variable has a value of 1 (one) if both teams have scored goals, and

Table 4
Features of each team used for the BTTS problem. The features are categorized as Performance Features, Sequential Features, and Market Features.

Performance	Sequential	Market
Matches Played (at Home)	p_{yy} - For scoring goals (at Home)	BTTS Yes Prob. (AVG)
Wins AVG (at Home)	p_{yn} - For scoring goals (at Home)	BTTS No Prob. (AVG)
Draws AVG (at Home)	p_{ny} - For scoring goals (at Home)	ML Home Prob. (AVG)
Loses AVG (at Home)	p_{nn} - For scoring goals (at Home)	ML Away Prob. (AVG)
Goals For AVG (at Home)	p_{yy} - For conceding goals (at Home)	ML Draw Prob. (AVG)
Goals Against AVG (at Home)	p_{yn} - For conceding goals (at Home)	BTTS Yes Prob. (MAX)
Points AVG (at Home)	p_{ny} - For conceding goals (at Home)	BTTS No Prob. (MAX)
Matches Scoring Goals AVG (at Home)	p_{nn} - For conceding goals (at Home)	ML Home Prob. (MAX)
Matches Conced. Goals AVG (at Home)	p_{yy} - For scoring goals (at Away)	ML Away Prob. (MAX)
Matches Played (at Away)	p_{yn} - For scoring goals (at Away)	ML Draw Prob. (MAX)
Wins AVG (at Away)	p_{ny} - For scoring goals (at Away)	BTTS Yes Prob. (MIN)
Draws AVG (at Away)	p_{nn} - For scoring goals (at Away)	BTTS No Prob. (MIN)
Loses AVG (at Away)	p_{yy} - For conceding goals (at Away)	ML Home Prob. (MIN)
Goals For AVG (at Away)	p_{yn} - For conceding goals (at Away)	ML Away Prob. (MIN)
Goals Against AVG (at Away)	p_{ny} - For conceding goals (at Away)	ML Draw Prob. (MIN)
Points AVG (at Away)	p_{nn} - For conceding goals (at Away)	BTTS Yes Prob. (STD)
Matches Scoring Goals AVG (at Away)	Current State (at Home)	BTTS No Prob. (STD)
Matches Conced. Goals AVG (at Away)	Current State (at Away)	ML Home Prob. (STD)
		ML Away Prob. (STD)
		ML Draw Prob. (STD)

0 (zero) otherwise. This feature is denoted by *btts*. More formally, *gh* and *ga* denote the number of goals scored by the home team and the away team, respectively. We define *btts* as follows:

$$btts = \begin{cases} 1, & \text{if } gh > 0 \text{ and } ga > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

4.3. Performance features

In football, the performance of a team in previous matches is an essential factor for determining its performance in the future. Thus, we extracted the overall standings for each team after each round of each season for all championships. As a result, we obtained the performance features for all teams after each round. Given that team performance may vary when playing at home or away, we also extracted the distinct standings, taking into account the location of the matches. This task resulted in the creation of two standing tables: performance at home and away. We call this group of features Performance Features. Since many prediction algorithms work better with normalized or standardized data, we divided the value of each feature of this group by the number of matches played. Thus, we transform the feature values into means, such as the average number of goals scored per game, the average number of points, and so on (see Table 4).

In short, this set of features represents the offensive and defensive strengths (at home and away) of teams during a season.

4.4. Sequential features

Although the performance features capture relevant knowledge about the past performance of teams, they do not include information about the sequence of events. To consider this kind of information, we exploit Markov chains, which are well known for modeling sequential data. We take into account two crucial scenarios: (i) matches in which one of the teams scored at least one goal, and (ii) games in which one side suffered at least one goal. For example, the fact that a team scores in a match might influence its performance in the next game, e.g., by making it more confident. Conversely, if a team does not score in a particular game, it may be less optimistic in the next match, which may affect its chances of scoring. For each team, we have built a Markov chain for modeling each one of these scenarios. In what follows, we describe these Markov chains for the case in which one team scores at least one goal in a match (it is analogous for the other scenarios).

To begin with, let $S = \{y, n\}$ be the state space of the process, i.e. the set of possible events, where y denotes that the team scored at least one goal in a match, and n otherwise. The process starts in one of these states (the result of the first match) and moves continuously from one state to another (the subsequent matches). If the Markov chain is currently in state s_i, then it moves to state s_j at the next step with probability p_{ij}. Thus, the probability p_{ij} corresponds to the transition probability and is defined as:

$$p_{ij} = P(X_n = j \mid X_{n-1} = i) \quad i, j \in S. \quad (2)$$

Therefore, for each championship, we have learned, for each team, two Markov chains from the data: one for home matches and another for away matches. All models have the same format (see Fig. 2), in which:

- the states correspond to the possible events (Yes, No);
- p_{yy} is the probability of the team scoring at least one goal in the next match given that it has scored in the previous match;
- p_{yn} is the probability of the team not scoring a goal in the next match given that it has scored in the previous match;
- p_{nn} is the probability of the team not scoring a goal in the next match given that it has not scored in the previous match;

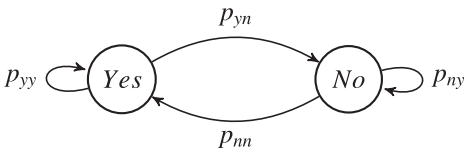

Fig. 2. Markov chain for the BTTS prediction problem.

- p_{ny} is the probability of the team scoring at least one goal in the next match, given that it has not scored in the previous match.

The Markov chains adjust the probabilities of each transition every time a new match is observed. We use the probabilities of these transitions as features. We refer to this set of features as Sequential Features (see Table 4).

4.5. Market features

A sports betting market represents, to some extent, the wisdom of the crowd. Hence, it may provide important signals for prediction models. First, given that odds can be thought of as probabilities, we use Shin's Method (Shin, 1993) to transform odds into probabilities in the Odds dataset. Next, to represent the general "opinion" of the market regarding a certain match, we calculate the average probabilities of all bookmakers. Even so, as this "opinion" is formed by several bookmakers, the averages may not represent the disagreements between bookmakers. Therefore, besides the average (avg), we computed the highest actual probability (max), the lowest actual probability (min), and the standard deviation (std) among all available probabilities. Formally, let $A = \{a_1, a_2, \ldots, a_n\}$ be the set of actual probabilities for a possible result, obtained from n bookmakers. We extracted $max(A)$, $min(A)$, $avg(A)$, and $std(A)$.

Finally, we can extract the favorites of the market by comparing the averages of the odds offered by all bookmakers, for Yes and No. Thus, the favorite class is the one with the lowest average odds.

Formally, for a match m, consider avg_yes_m and avg_no_m as the average odds of bookmakers for the Yes and No classes, respectively. Denote $fav(m)$ as the function that defines the favorite of the market:

$$fav(m) = \begin{cases} 1, & \text{if } avg\_yes_m \leq avg\_no_m \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

These features are referred to as Market Features (see Table 4).

5. Predictive models

In this section, we detail the classification models evaluated in this paper.

5.1. Majority Class Classifier (MJC)

A naive approach to predicting whether both teams will score in a match involves identifying the overall distribution of results for the corresponding championship. If most of the matches in the training set end with both teams scoring, the classifier predicts Yes; otherwise, it predicts No. In this paper, we refer to this model as the Majority Class Classifier (MJC). To determine the majority class, we basically analyze the distribution of BTTS results (Yes or No). Fig. 3 shows the distributions of "BTTS: Yes" considering all championships (left-hand side) and per championship (right-hand side).

Considering all championships, we notice that 50.61% of the matches finished with both teams scoring. Although there is a small bias toward Yes, the difference among classes is not statistically significant (Table A.6 of the Appendix presents the t-test values).

Regarding the distributions per championship, we observe that in the championships Italy A, Germany A, and Netherlands A, the majority class is Yes (both teams score), while in Brazil A and Portugal A, the majority class is No. For the championships France A, England A, Spain A, and Brazil B, we do not notice any pronounced bias toward one class or the other, which is confirmed by the high observed p-values. However, in practice, the classifier will follow the majority class regardless of the significance of the test. That is, for France A and England A, it predicts No, and for Spain A and Brazil B, it predicts Yes.

These observed class imbalances, although subtle, provide some evidence that a model that strictly follows the majority class (also called the Zero Rule method) may provide better predictions than a random one. For our analyses, this will serve as a lower-bound baseline. That is, we expect that more sophisticated models should be able to outperform it.

5.2. Poisson-Based classifiers

We used three classifiers supported by Poisson models as baselines. These models have been widely used to predict the number of goals in football matches and have served as a good baseline for several related works, such as Boshnakov et al. (2017), Koopman and Lit (2015), Koopman and Lit (2019) and Owen (2011). Although not designed for the BTTS prediction problem, they can be easily adapted for this task, as we point out in Section 3.

The basic model that served as the basis for so many others is the model proposed by Maher (1982). The central premise of this model is that the number of goals predicted by the home team and away team in a particular match are independent Poisson variables, whose means are determined by the respective attack and defense qualities of each team. Formally, considering h the home team, v the visitor team, and X and Y the number of goals scored by each one, respectively, then:

$$X \sim \text{Poisson}(\alpha_h\ \beta_v\ \gamma)$$

$$Y \sim \text{Poisson}(\alpha_v\ \beta_h)$$

in which X and Y are independent, $\alpha_i, \beta_i > 0$, $\forall i \in (h, v)$, α_i measures the attack strength, β_i measures the defense strength, and $\gamma > 0$ is a parameter that refers to home

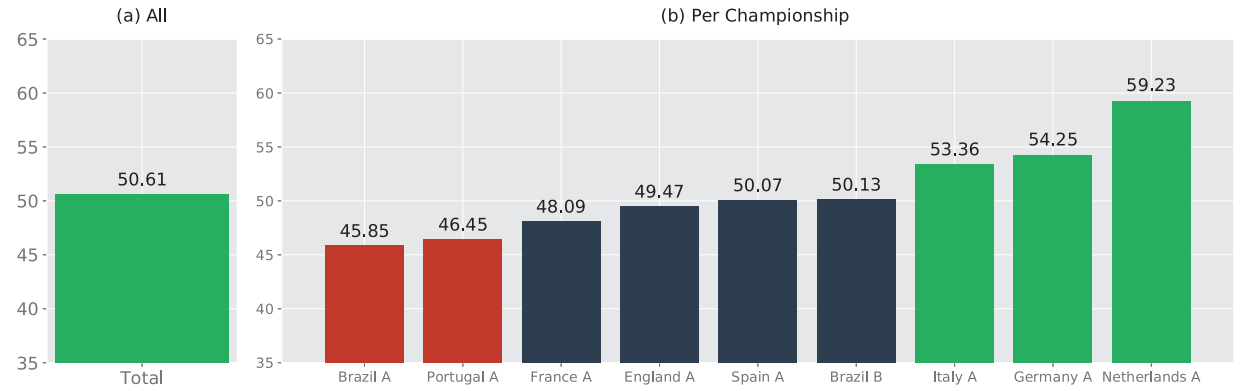

Fig. 3. Distribution of "BTTS: Yes" considering all championships combined (a) and per championship (b).

advantage.[1] In this paper, we refer to this model as the Independent Poisson Classifier (IPO).

Dixon and Coles (1997) introduced an improvement to the independent goal model proposed by Maher (1982). They showed that the model was not able to predict low-scoring matches and introduced a dependence parameter for the following match outcomes: $0-0$, $1-0$, $0-1$, and $1-1$. They also proposed two weighting functions to give more importance to the most recent matches. In this paper, we call this model the Dixon and Coles Classifier (DAC).

Rue and Salvesen (2000) extended the framework of Dixon and Coles (1997) and developed a Bayesian dynamic generalized linear model to update the prior estimates of parameters, which are analyzed in continuous time by Markov chain Monte Carlo methods. In this paper, we refer to this model as the Rue and Salvesen Classifier (RAS).

For all cases, we extract a matrix $S \in \mathbb{R}^{a \times b}$ from the models, in which the indexes a, b are the possible number of goals for the home team and away team, and the values are the probabilities of each possible match result. Thus, the prediction of the classifiers for the BTTS problem can be defined as follows:

$$P_{no} = S[0,0] + S[0,1:] + S[1:,0]$$

$$P_{yes} = 1 - P_{no}$$

5.3. Machine learning classifiers

We evaluated three ML classifiers: Gaussian Naive Bayes (GNB), Logistic Regression (LRE), and Gradient Boosting (XGB). These are strong classifiers typically used in a wide range of complex prediction domains, including football analytics, as detailed in Section 3.

5.4. Market classifiers

We define two classifiers based on betting market data: the Average Market Classifier (AMK) and the Fairest Bookmaker Classifier (FBO).

AMK makes predictions based on the average opinion of 19 bookmakers. That is, it defines the probabilities for Yes and No according to average odds published by the bookmakers, normalized by Shin's model (Štrumbelj, 2014). It should be a strong predictor under the market efficiency hypothesis.

To build the FBO classifier, we analyzed which bookmaker had the lowest average booksum in the training set. A bookmaker with a low booksum is expected to be a better place to bet. In this way, making comparisons against it can allow us to identify profit opportunities (see further details on this in Section 6.2).

Fig. 4 shows that the bookmaker 1xBet has the lowest booksum. That is, 1xBet seems to offer the fairest odds compared to other bookmakers. In this way, FBO makes predictions based strictly on the odds from 1xBet.

6. Evaluation

In this section, we describe several experiments we conducted to answer the research questions stated in Section 1. For answering RQ1 and RQ2, we take a closer look at the performance of the classifiers defined in the previous section. For answering RQ3, we analyze the performance of different betting strategies based on the predictions of the classifiers.

6.1. Setup and evaluation protocol

We split our dataset into 66.67% of the instances for training and the remaining 33.33% for testing. The training set contained matches from the first four seasons (2013–2016), while the test set contained matches from the last two seasons (2017–2018), from the fifth round. For all cases, we evaluated the classifiers using the landmark window strategy (growing window) (Ryan et al., 2013). Specifically, in each iteration, the test instances of the previous iteration are added as training instances and used for predicting the next time window. In our case, after each iteration, the matches predicted are added as training for predicting the next round. Thus, we retrain the models after each round.

[1] Home advantage is a factor that represents an additional strength for the team playing at home.

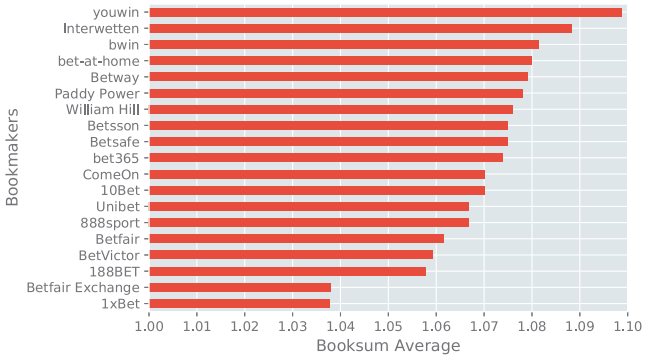

Fig. 4. Average booksum of bookmakers.

For building the Poisson-based classifiers (PBCs), IPO, DAC, and RAS, we used the *goalmodel* package.[2] For GNB and LRE we used *scikit-learn*[3] and for XGB we used *XGBoost*[4] as ML libraries. We tuned the learning parameters of all algorithms by means of five-fold cross-validation.

In the training phase, we had different setups. The PBCs only need the scoreboards of the matches. For prediction purposes, it is useful to weigh the influence of each match result, so that matches far back in time have less influence than the more recent ones. We searched for the optimal value for this time-weighting parameter ξ, considering the number of days as the time unit. To this end, we evaluated values between 0 and 0.008 (similar to Boshnakov et al. (2017)). For example, the optimal value found for the model that predicts the first round of the 2017 season was $\xi = 0.001$.

While PBCs only need the past scoreboards, the ML models can be fed with different sets of features. Thus, for the ML classifiers, we performed an ablation study with three experiments in order to evaluate which groups of features performed best.

In the first experiment, we used only features derived from the Match dataset, i.e. *Performance Features* and *Sequential Features*. In the second experiment, we used only features derived from the Odds dataset, i.e. *Market Features*. Finally, in the third experiment, we used all available features. To identify each classifier individually, we added to a label to the classifier, an underscore followed by a capital letter that identifies the group of features used. For example, for the GNB classifier we have: GNB_P (Performance and Sequential Features), GNB_M (Market Features), and GNB_A (all available features). We used the same logic for LRE and XGB.

For RQ1 and RQ2, we evaluated the models in terms of accuracy (ACC) and Brier score (BRS). Accuracy is the percentage of correct predictions, while the Brier score measures model performance in forecasting the probability of each class. We express all the probabilities as percentages. Formally, considering N the number of matches, m an instance of a game, $\hat{y}_m$ the probability forecast by the model for BTTS, and o_m the actual outcome (0 if it does not happen; 1, otherwise), we can define the Brier score as:

$$BRS = \frac{1}{N}\sum_{m=1}^{N}(\hat{y}_m - o_m) \tag{4}$$

For RQ3, we evaluated the models in terms of profitability (PRF) and return on investment (ROI). Profitability measures how much money we may earn or loose (balance) after betting based on a given strategy. In other words, when we place a bet on a match, if we correctly predict the result, we will receive a profit based on the odds offered by the bookmaker. On the other hand, if we fail to predict the outcome, we will lose all the wagered amount. Thus, we define profitability as the sum of these profits and losses, considering a group of matches. ROI measures the gain or loss generated on an investment relative to the amount of money invested.

6.2. Results and discussion

In this section, we revisit the research questions of this paper, providing answers and discussions based on the attained results.

RQ1: How difficult is the BTTS prediction problem? To answer this question, we can look at the accuracy of two empirical classifiers: AMK and MJC.

From the perspective of pure market efficiency, it should not be possible to have a better predictor than the market's "opinion" itself. Thus, in some sense, AMK may be regarded as an upper bound. That is, it should not be possible to devise a predictive model better than the market. Thus, comparing AMK (the best we can get under a hypothesis of market efficiency) with MJC, (the simplest baseline we can think of) might help us to understand the difficulty of the problem.

The results show that, considering all matches in the test set, AMK has an accuracy of 55.34% against 51.51% with MJC (see Fig. 5). Knowing that BTTS is a binary classification problem, we can observe that the accuracies of these classifiers are only a little better than a random classifier, which would have an average accuracy of 50%. Besides, the difference between AMK and MJC is small (less than 4%), which indicates that there might not be much room for improvement. Thus, these results provide

[2] https://github.com/opisthokonta/goalmodel.

[3] https://scikit-learn.org/stable/.

[4] https://xgboost.readthedocs.io/en/latest/.

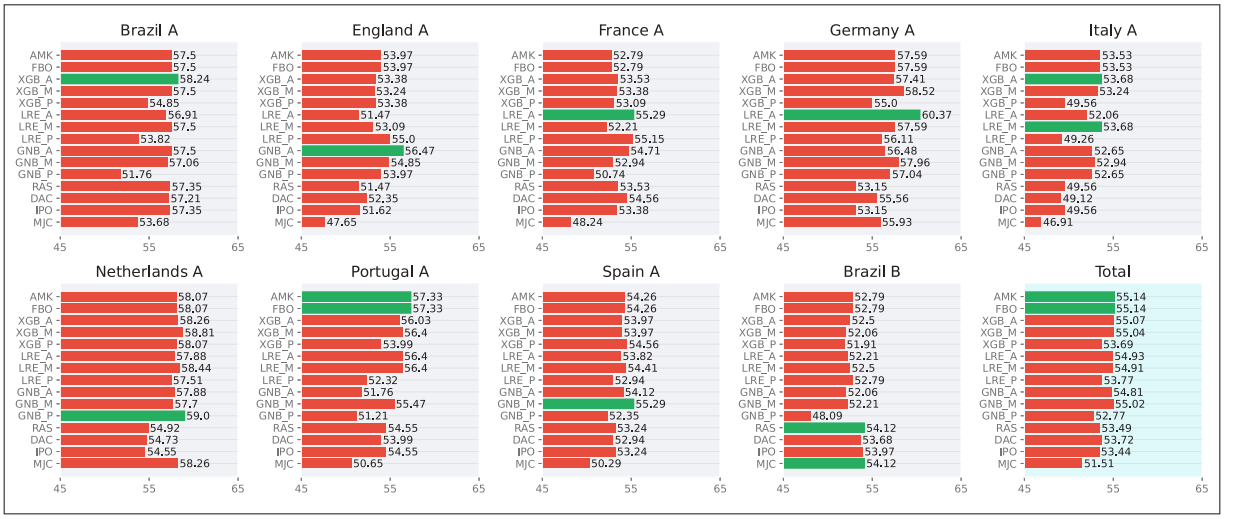

Fig. 5. Classifier performance in terms of accuracy.

good evidence that BTTS may be regarded as a problem that is hard to predict.

We can make the same analysis considering matches grouped by championships (see Fig. 5). In this scenario, through Wilcoxon tests (see Table A.7 of the Appendix), we observed that AMK is not significant better than MJC in the championships Germany A, Netherlands A, and Spain A. In Brazil B specifically, AMK presented even worse performance than MJC. For the other championships, AMK showed significantly higher performance. These results reinforce the difficulty of the problem, where in some cases the market is not even better than an approach as simple as MJC. These results also hint that some championships may be easier to predict than others.

RQ2: Are ML classifiers capable of predicting BTTS better than bookmakers?

This research question aims to evaluate whether classifiers based on carefully engineered features are able to make more accurate predictions than the market, represented in this paper by the classifiers AMK and FBO. Fig. 5 presents the performance of classifiers in terms of accuracy, while Fig. 6 depicts it in terms of the Brier score.

First, we analyze the classifiers when they are fed only with data extracted from the performance of the teams, i.e. when they do not use market data. In this scenario, considering all championships, Poisson-based classifiers (PBCs) and machine learning classifiers (MLCs) had similar performances, and only GNB was unable to perform better than MJC. On the other hand, none of them was able to beat the bookmakers.

A second view of the problem is to assess how the classifiers behave when fed with market information. Analyzing the ML classifiers when fed with market features, we can observe that they improved their performance. In this case, the difference between the accuracy of the best classifiers was quite small: XGB_M (ACC: 55.04, BRS: 0.2462), LRE_M (ACC: 54.91, BRS: 0.2459), and GNB_M (ACC: 55.02, BRS: 0.2596). And there was no statistical difference between these classifiers and AMK (ACC: 54.91, BRS: 0.2459). One hypothesis for this insignificant difference is that, if the market is efficient when the classifiers are fed only market information, they end up reproducing the same predictive behavior of the market. To assess whether this is true, we calculated the Kappa coefficient between the ML classifiers and the market classifiers. This coefficient can be used to describe the level of agreement between classifiers. The closer to 1, the stronger the agreement between them. Fig. 7 shows the coefficient values.

The results confirm that the ML classifiers when fed market features have nearly perfect correlation with AMK ($k > 0.8$), whereas when they are fed performance features they have moderate ($0.4 < k < 0.6$) or fair ($0.2 < k < 0.4$) correlation. In future work, we plan to investigate further the characteristics of the matches in which the classifiers diverged from the market in order to find eventual patterns of market inefficiency.

Lastly, when we used all available features, the ML classifiers showed no significant improvement over the results of the previous experiments, which probably means that the market features already bear the strongest signals and thus seems to incorporate all the information of the performance features.

Finally, we can answer the research question from two perspectives. Given that the difference between the best ML classifiers and market classifiers is minimal, we can assume their performances as being equivalent. In this sense, the answer to RQ2 is negative. However, in some championships, the ML classifiers are as good as AMK and slightly better than FBO, in terms of the Brier score. This provides hints that ML classifiers have a better calibration in predicting the probabilities of BTTS than FBO, which could allow us to create profitable strategies against the 1xBet bookmaker. Thus, in this perspective, considering only the worst bookmaker, the answer to RQ2 is positive.

RQ3: Are ML classifiers useful for devising profitable betting strategies?

This research question aims at evaluating whether the ML classifiers are useful for devising profitable systematic

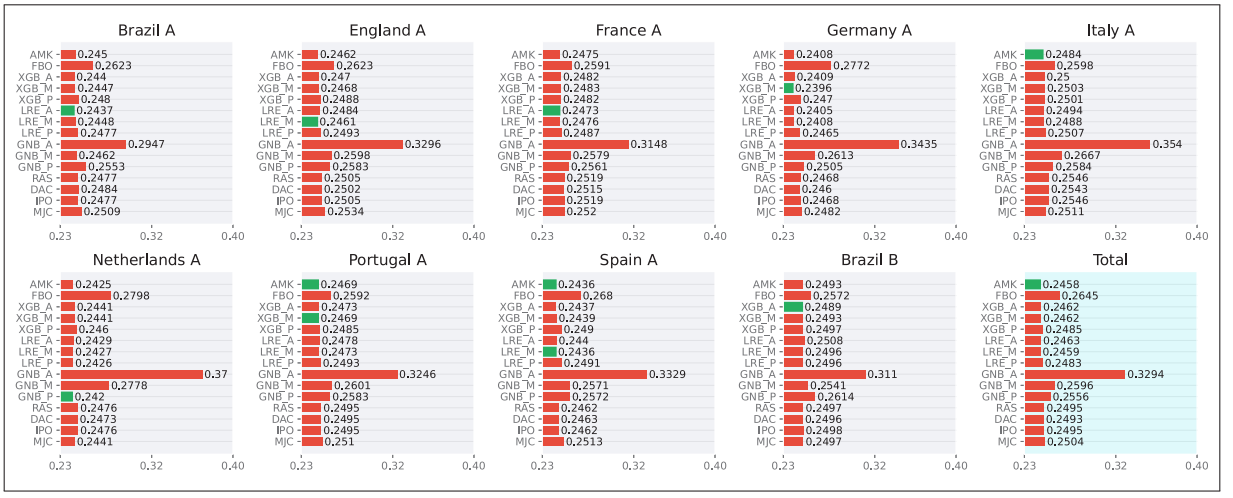

Fig. 6. Classifier performance in terms of the Brier score (the smaller the better).

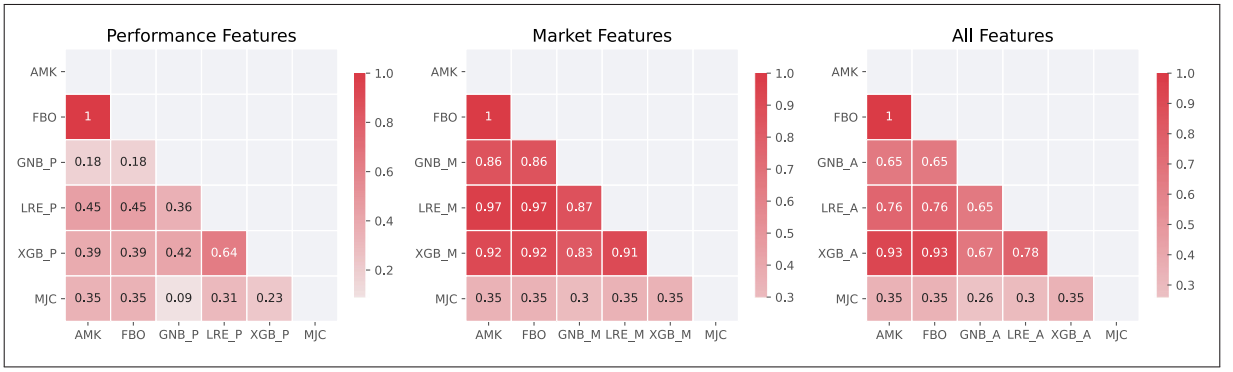

Fig. 7. Kappa coefficient of the correlation of classifier predictions.

strategies for the betting market. To this end, we consider three betting strategies: naive betting (NB), expected value (EV), and proportional betting (PB). We evaluated the use of these three strategies for each of the classifiers in terms of profitability (PRF) and return on investment (ROI).

NB is a straightforward strategy adopted by less experienced bettors who end up betting on the result they think will happen, regardless of the odds offered by bookmakers. Thus, in this naive approach, we bet $1 on the result (Yes or No) predicted by a given classifier.

In the second strategy, we bet according to the expected value (EV). This strategy takes into account the difference between the probability of the market and the model. It is commonly used in related works, such as Boshnakov et al. (2017), Angelini and De Angelis (2017), and Koopman and Lit (2019). Formally, the expected value of a bet on a match is given by Eq. (5), where A represents a given result (Yes or No), $P(A)$ is the probability of the result A according to the model, and $odds(A)$ is the published odds for the result A. Thus, the strategy is to bet only on results for which the expected value is positive, $EV(A) > 0$.

$$EV(A) = P(A) * odds(A) - 1 \tag{5}$$

The third strategy is proportional betting (PB). PB takes into account the odds and the expected value in order to find the optimal stake on a bet given by Eq. (6). The PB strategy is similar to the Kelly criterion (Kelly, 2011), but it differs by not taking into account the overall wealth of the bettor and fixing the maximum stake at $1 per match, as done by Boshnakov et al. (2017).

$$PB(A) = \frac{(odds(A) + 1)P(A) - 1}{odds(A)} \tag{6}$$

Initially, we tested the strategies against the average odds offered by the market. As expected, in an efficient market, no classifier was capable of devising a profitable strategy. Fig. 8 shows that all classifiers presented negative profitability when considering all matches of the test set.

An alternative to making a profit is always trying to bet on the bookmaker that offers the best offer (the maximum odds) for each match, but for this situation, it would be necessary for the bettor to have accounts in 19 bookmakers, which may not sound reasonable. Thus, we evaluated the strategies against a single bookmaker. For that, we chose the bookmaker 1xBet, which generally offers better odds than other bookmakers, as shown in

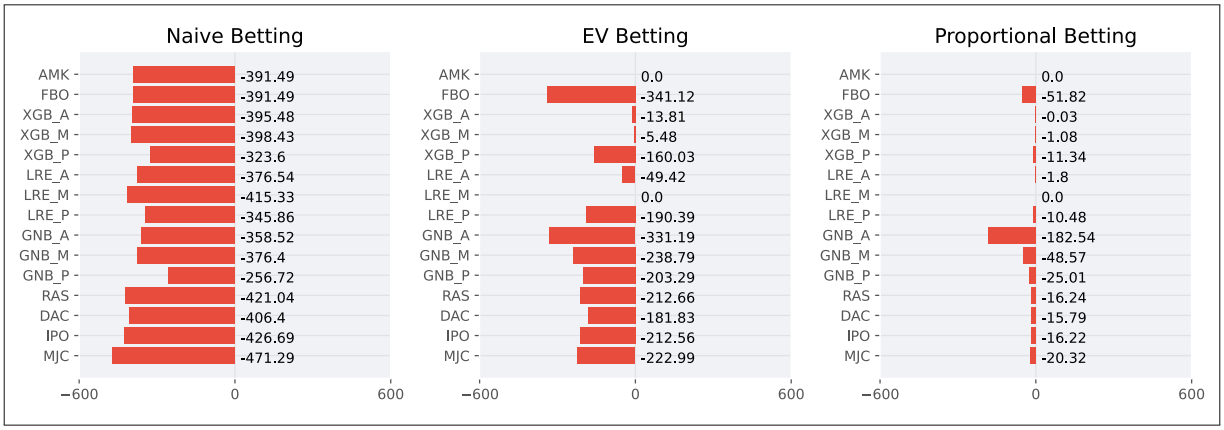

Fig. 8. Profitability obtained through different strategies (NB, EV, and PB) against average market odds.

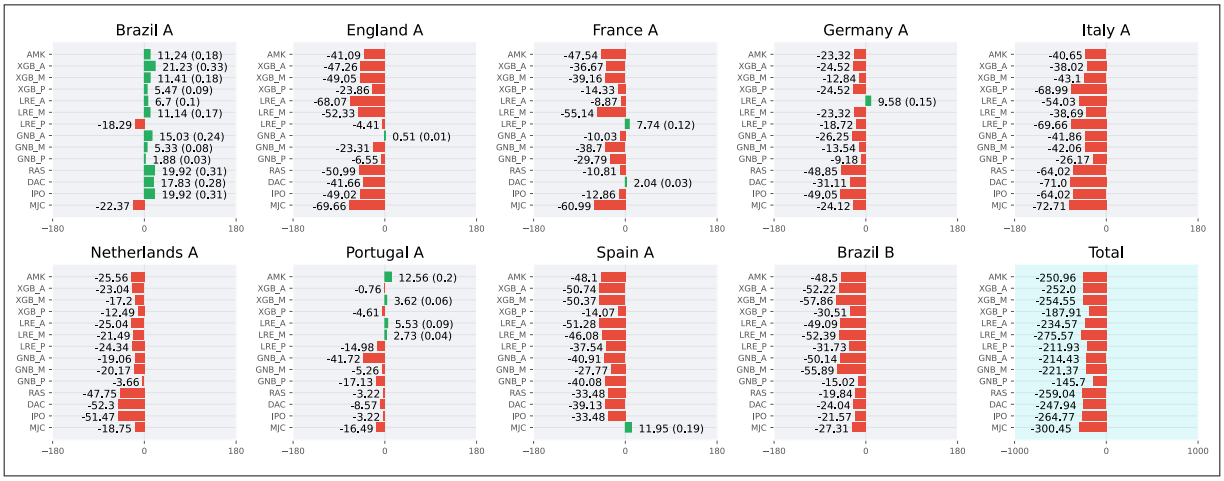

Fig. 9. Profitability (and ROI) obtained through the naive betting (NB) strategy against the bookmaker 1xBet.

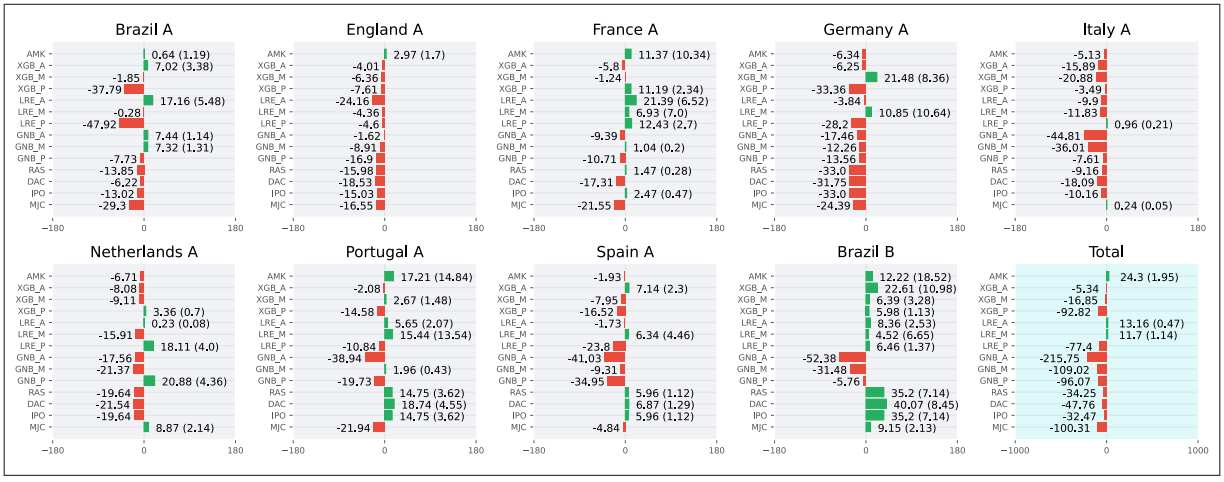

Fig. 10. Profitability (and ROI) obtained through the expected value (EV) strategy against the bookmaker 1xBet.

Section 5.4. Figs. 9, 10, and 11 present the profitability and ROI obtained through the strategies NB, EV, and PB, respectively (Tables A.8–A.10 of the Appendix present the Wilcoxon test values for the aforementioned strategies).

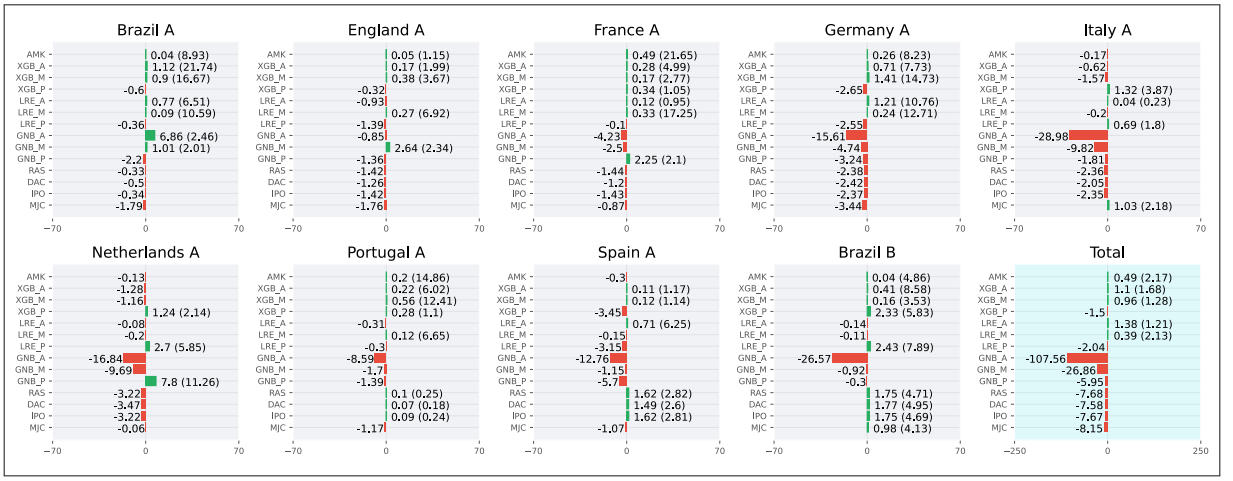

Fig. 11. Profitability (and ROI) obtained through the proportional betting (PB) strategy against the bookmaker 1xBet.

In general, no classifier was successful using the NB strategy. Only in the Brazil A championship was it the case that more than one classifier was capable of providing a positive return, although with an insignificant ROI (see Fig. 9).

Regarding the EV strategy, the results were remarkably better, but gains remained small. In general, the best classifier was the market itself (AMK), which achieved profitability of 24.3, but with an ROI of only 1.95%. Analyzing by championship (see Fig. 10), in England A and Italy A, ML classifiers were not efficient, but in the others, LRE and XGB were capable of obtaining profits in several matches, mainly in the championships France A, Brazil B, and Portugal A. The results show that there may be championships with more profit opportunities than others.

Regarding the PB strategy, compared to EV, the ROI was generally better, although profitability was worse. Due to the small difference between the probabilities predicted by the classifiers and those predicted by the bookmaker, the PB strategy ended up establishing a rather small stake. Thus, we would need a significant increase in the maximum value fixed per match for this strategy to reach a consistent profit.

So far, we have tested strategies considering betting on all matches. However, we could follow a strategy of evaluating a bet on only a specific group of matches. To further analyze this idea, we re-evaluated the profitability of classifiers considering only a group of matches in which the probability of "BTTS: Yes", according to the market, belongs to a given range.

We evaluated the profitability in all possible probability ranges (considering integer numbers) using the three strategies against the bookmaker 1xBet. The results demonstrated that the EV strategy was the most effective for all classifiers. Fig. 12 displays the results for all possible intervals. Table 5 displays the best results, highlighting the classifier, the range of probabilities (for Yes) that offered the best return, the number of matches identified by the classifier with positive EV (within range), and the PRF and ROI obtained with the bets.

Table 5
Results of classifiers using the EV strategy in a given group of matches.

Classifier	Prob. Range	# Matches	PRF	ROI
MJC	39%–42%	303	33.68	11.12%
IPO	32%–51%	2381	47.50	1.99%
DAC	32%–44%	661	51.53	7.8%
RAS	32%–51%	2380	44.72	1.88%
GNB_P	39%–42%	304	31.71	10.43%
GNB_M	43%–48%	1409	39.48	2.80%
GNB_A	50%–51%	277	28.21	10.18%
LRE_P	50%–63%	1904	61.97	3.25%
LRE_M	33%–48%	432	43.68	10.11%
LRE_A	30%–52%	1732	36.14	2.09%
XGB_P	55%–63%	1138	40.62	3.57%
XGB_M	55%–64%	760	31.29	4.12%
XGB_A	55%–64%	806	51.85	6.43%
AMK	39%–49%	350	38.32	13.53%

These results are promising, since all classifiers achieved some positive returns. Even GNB_A and MJC, which had not obtained good results previously, were capable of making a profit. We can observe that these classifiers were capable of obtaining profit in matches in which the probability for Yes is considered small for the bookmaker (from 39%).

LRE_M achieved the highest overall profit (61.57), but for that, it had to bet on many matches, which resulted in a small ROI (3.25%) compared to the others. On the other hand, AMK presented the highest ROI (13.53%), but modest profitability (38.32). The GNB and XGB classifiers also had good results, but curiously in different probability ranges. While the GNB classifiers were better when the probability for Yes was between 39% and 51%, XGB classifiers performed better when this probability was between 55% and 64%. Among Poisson-based classifiers, DAC achieved the best result. These results hint that distinct classifiers can be more profitable in different matches. In future work, we will focus on understanding these patterns and creating more sophisticated and smarter strategies for identifying more profitable groups of matches.

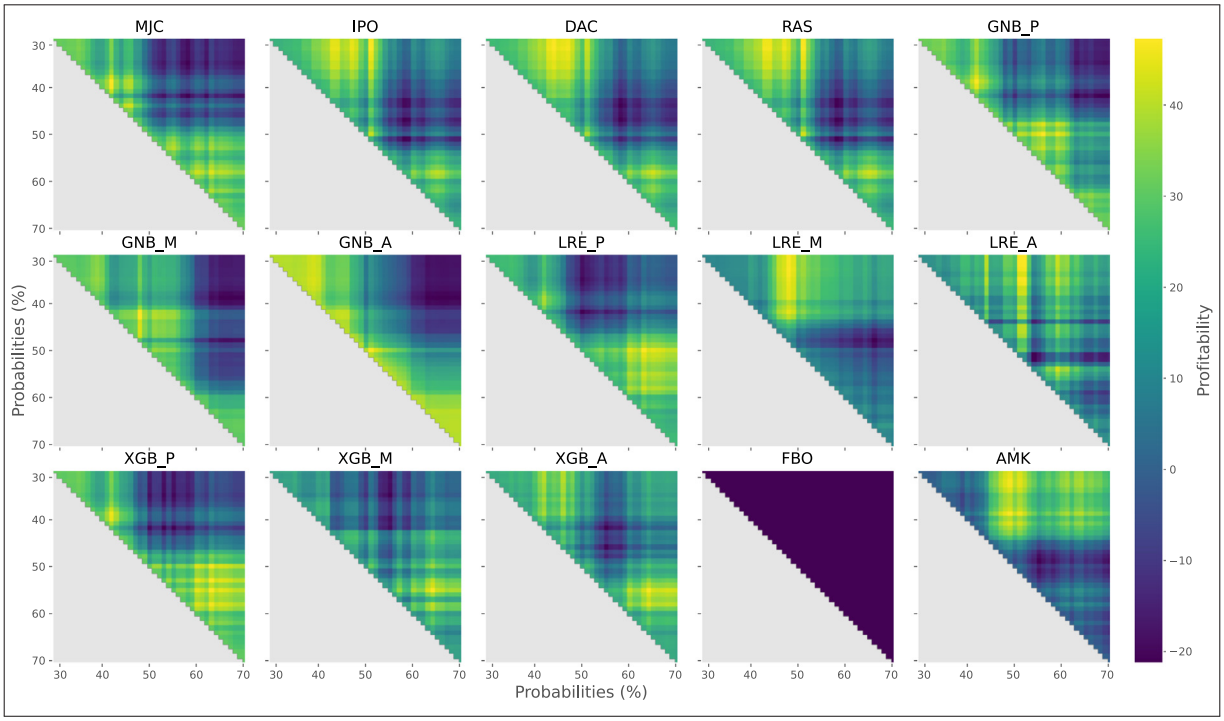

Fig. 12. Profitability obtained through the expected value (EV) strategy against the bookmaker 1xBet considering a range of probabilities. The *y*-coordinate is the beginning of the interval of probabilities, and the *x*-coordinate is the end. Profitability obtained in each range is represented by the variation of colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Finally, as in RQ2, we can also answer RQ3 from two perspectives. When the classifiers bet on all matches against the average market odds, they are not able to obtain profit from the strategies evaluated in this work. Thus, in this sense, the answer to RQ3 is negative. However, when we choose a fairer bookmaker and limit the number of bets, whether by championship or by a range of probabilities, the classifiers can be useful for devising profitable betting strategies. From this perspective, the answer to RQ3 is positive.

6.3. Reproducibility

We provide all of the code and datasets for this work on the Github website of the first author.[5] We also provide the code for reconstructing our training data (feature engineering), as well as additional information about the results of the experiments and statistical tests.

7. Conclusion and future work

In this paper, we investigated a hitherto unexplored research question in the field of football forecasting, viz. the "both teams to score" (BTTS) prediction problem. We evaluated the performance of machine learning model predictions in terms of accuracy as well as efficiency in the betting market. For this, we put together a large-scale dataset comprising nine national leagues and 19 bookmakers.

Table A.6
t-Test for comparing class distributions of BTTS results, $H_0 : yes = no \mid H_1 : yes \neq no$ (relative to Figure Fig. 3).

Championship	*t*-statistic	*p*-value
Brazil A	−3.24409	0.00120
Brazil B	0.10256	0.91832
England A	−0.41028	0.68166
France A	−1.48826	0.13689
Germany A	2.98221	0.00292
Italy A	2.62129	0.00885
Netherlands A	6.57011	0.00000
Portugal A	−2.41681	0.01581
Spain A	0.05128	0.95911
All	1.383128	0.16664

Table A.7
Wilcoxon test to evaluate whether AMK is better than MJC in terms of accuracy.

Championship	*t*-statistic	*p*-value
Brazil A	3175.0	0.020544
Brazil B	546.0	0.207578
England A	20043.0	0.013810
France A	20100.0	0.073009
Germany A	2064.0	0.355809
Italy A	22518.0	0.012284
Netherlands A	650.0	0.888638
Portugal A	11203.5	0.017608
Spain A	35295.0	0.171014
All	776985.0	0.000002

[5] https://github.com/igormago/btts_ijof.

Table A.8
Wilcoxon test to compare the profitability of the classifiers following the NB strategy against 1xBet.

	IPO	DAC	RAS	GNB_P	GNB_M	GNB_A	LRE_P	LRE_M	LRE_A	XGB_P	XGB_M	XGB_A	FBO	AMK
IPO	–	0.1468	0.1763	0.0073	0.5488	0.7697	0.5013	0.1156	0.4059	0.2132	0.1985	0.2037	0.1932	0.1932
DAC	0.1468	–	0.0964	0.009	0.4925	0.7001	0.56	0.095	0.36	0.2492	0.1696	0.1721	0.1625	0.1625
RAS	0.1763	0.0964	–	0.0087	0.5057	0.7204	0.5437	0.101	0.3695	0.2376	0.1763	0.1811	0.1715	0.1715
GNB_P	0.0073	0.009	0.0087	–	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GNB_M	0.5488	0.4925	0.5057	0.0	–	0.0	0.0	0.0	0.7205	0.0	0.0	0.0	0.0	0.0
GNB_A	0.7697	0.7001	0.7204	0.0	0.0	–	0.0	0.0	0.0002	0.0	0.0	0.0	0.0	0.0
LRE_P	0.5013	0.56	0.5437	0.0	0.0	0.0	–	0.0	0.0	0.0159	0.0	0.0	0.0	0.0
LRE_M	0.1156	0.095	0.101	0.0	0.0	0.0	0.0	–	0.0	0.0	0.0	0.0	0.3014	0.3014
LRE_A	0.4059	0.36	0.3695	0.0	0.7205	0.0002	0.0	0.0	–	0.0	0.0	0.0	0.0	0.0
XGB_P	0.2132	0.2492	0.2376	0.0	0.0	0.0	0.0159	0.0	0.0	–	0.0	0.0	0.0	0.0
XGB_M	0.1985	0.1696	0.1763	0.0	0.0	0.0	0.0	0.0	0.0	0.0	–	0.529	0.0001	0.0001
XGB_A	0.2037	0.1721	0.1811	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.529	–	0.0007	0.0007
FBO	0.1932	0.1625	0.1715	0.0	0.0	0.0	0.0	0.3014	0.0	0.0	0.0001	0.0007	–	–
AMK	0.1932	0.1625	0.1715	0.0	0.0	0.0	0.0	0.3014	0.0	0.0	0.0001	0.0007	–	–

Table A.9
Wilcoxon test to compare the profitability of the classifiers following the EV strategy against 1xBet.

	IPO	DAC	RAS	GNB_P	GNB_M	GNB_A	LRE_P	LRE_M	LRE_A	XGB_P	XGB_M	XGB_A	FBO	AMK
IPO	–	0.1395	0.7463	0.3991	0.2042	0.0018	0.8608	0.0001	0.0719	0.8631	0.0979	0.0614	0.8091	0.0002
DAC	0.1395	–	0.1518	0.4932	0.2563	0.0031	0.7078	0.0	0.0509	0.7199	0.0509	0.0249	0.7053	0.0
RAS	0.7463	0.1518	–	0.411	0.211	0.002	0.8432	0.0001	0.0682	0.8465	0.0912	0.0574	0.7956	0.0002
GNB_P	0.3991	0.4932	0.411	–	0.0	0.0	0.0011	0.0	0.0007	0.0002	0.0041	0.0002	0.0	0.0
GNB_M	0.2042	0.2563	0.211	0.0	–	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GNB_A	0.0018	0.0031	0.002	0.0	0.0	–	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LRE_P	0.8608	0.7078	0.8432	0.0011	0.0	0.0	–	0.0	0.0002	0.0031	0.0099	0.0042	0.0078	0.0
LRE_M	0.0001	0.0	0.0001	0.0	0.0	0.0	0.0	–	0.0	0.0	0.0684	0.0	0.0	0.0
LRE_A	0.0719	0.0509	0.0682	0.0007	0.0	0.0	0.0002	0.0	–	0.0	0.0054	0.0	0.0	0.0
XGB_P	0.8631	0.7199	0.8465	0.0002	0.0	0.0	0.0031	0.0	0.0	–	0.0	0.0	0.2443	0.0
XGB_M	0.0979	0.0509	0.0912	0.0041	0.0	0.0	0.0099	0.0684	0.0054	0.0	–	0.0025	0.0	0.5753
XGB_A	0.0614	0.0249	0.0574	0.0002	0.0	0.0	0.0042	0.0	0.0	0.0	0.0025	–	0.0	0.0012
FBO	0.8091	0.7053	0.7956	0.0	0.0	0.0	0.0078	0.0	0.0	0.2443	0.0	0.0	–	0.0
AMK	0.0002	0.0	0.0002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5753	0.0012	0.0	–

Table A.10
Wilcoxon test to compare the profitability of the classifiers following the PB strategy against 1xBet.

	IPO	DAC	RAS	GNB_P	GNB_M	GNB_A	LRE_P	LRE_M	LRE_A	XGB_P	XGB_M	XGB_A	FBO	AMK
IPO	–	0.0639	0.2613	0.6457	0.0035	0.144	0.4275	0.1641	0.1212	0.2726	0.1157	0.1275	0.0	0.1716
DAC	0.0639	–	0.065	0.6152	0.0034	0.1389	0.3318	0.2269	0.1613	0.1988	0.1679	0.1894	0.0	0.2346
RAS	0.2613	0.065	–	0.6453	0.0035	0.1441	0.4282	0.1644	0.1213	0.273	0.1156	0.1277	0.0	0.1716
GNB_P	0.6457	0.6152	0.6453	–	0.0469	0.1016	0.5941	0.1977	0.2177	0.9279	0.2027	0.2396	0.0	0.217
GNB_M	0.0035	0.0034	0.0035	0.0469	–	0.0089	0.0	0.0001	0.0001	0.0	0.0001	0.0001	0.0	0.0001
GNB_A	0.144	0.1389	0.1441	0.1016	0.0089	–	0.2005	0.0223	0.0355	0.2014	0.0201	0.0248	0.4742	0.0248
LRE_P	0.4275	0.3318	0.4282	0.5941	0.0	0.2005	–	0.0	0.0	0.1138	0.0	0.0	0.0	0.0
LRE_M	0.1641	0.2269	0.1644	0.1977	0.0001	0.0223	0.0	–	0.6385	0.0	0.5908	0.0383	0.0	0.0357
LRE_A	0.1212	0.1613	0.1213	0.2177	0.0001	0.0355	0.0	0.6385	–	0.0	0.3574	0.1256	0.0	0.2748
XGB_P	0.2726	0.1988	0.273	0.9279	0.0	0.2014	0.1138	0.0	0.0	–	0.0	0.0	0.0	0.0
XGB_M	0.1157	0.1679	0.1156	0.2027	0.0001	0.0201	0.0	0.5908	0.3574	0.0	–	0.0174	0.0	0.9018
XGB_A	0.1275	0.1894	0.1277	0.2396	0.0001	0.0248	0.0	0.0383	0.1256	0.0	0.0174	–	0.0	0.2246
FBO	0.0	0.0	0.0	0.0	0.0	0.4742	0.0	0.0	0.0	0.0	0.0	0.0	–	0.0
AMK	0.1716	0.2346	0.1716	0.217	0.0001	0.0248	0.0	0.0357	0.2748	0.0	0.9018	0.2246	0.0	–

First, we described the characteristics of the domain and observed that the BTTS prediction problem is a difficult task, where even the market is only slightly better than a simple model that always predicts the majority class. Next, we performed a careful feature engineering process in order to build robust classifiers for this problem. We conducted a comprehensive set of experiments that showed that machine learning classifiers are capable of beating the market in terms of prediction accuracy, although only in a few scenarios. In fact, the best performing features were the ones extracted from the betting market itself, where the classifiers basically learned to mimic the market in most of the matches.

We also performed an experiment where we used the classifiers' predictions as input to betting strategies. The assumption was that if we could systemically beat the market in terms of profitability, then there would be indications that the market was inefficient. We observed that when betting on all matches, the classifiers did not reach a significant profit margin. On the other hand, when we selected specific championships or certain game sets, there were indications that the classifiers were more profitable, thus making machine learning approaches even more interesting for this kind of problem.

As future work, we plan to include additional features (e.g. the number of shots and corners, and the duration of ball possession), new feature engineering methods, classifiers based on deep learning, and approaches for finding more clever betting strategies through machine learning techniques.

Appendix. Statistical tests

See Tables A.6–A.10.

References

Anderson, C., & Sally, D. (2013). *The numbers game: Why everything you know about football is wrong*. Penguin UK.

Angelini, G., & De Angelis, L. (2017). PARX model for football match predictions. *Journal of Forecasting*, *36*(7), 795–807.

Angelini, G., & De Angelis, L. (2019). Efficiency of online football betting markets. *International Journal of Forecasting*, *35*(2), 712–721.

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741–755.

Berrar, D., Lopes, P., Davis, J., & Dubitzky, W. (2017). OSF | the 2017 soccer prediction challenge. https://osf.io/ftuva/. (Accessed 22 October 2018).

Berrar, D., Lopes, P., Davis, J., & Dubitzky, W. (2018a). Guest editorial: Special issue on machine learning for soccer. *Machine Learning*, *108*, 1–7.

Berrar, D., Lopes, P., & Dubitzky, W. (2018b). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 1–30.

Betsonly (2017). Most popular football betting markets | betsonly.com. http://www.betsonly.com/sport-betting/popular-football-betting-markets. (Accessed 20 June 2018).

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*(2), 458–466.

Bunker, R. P., & Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied Computing and Informatics*.

Constantinou, A. C. (2019). Dolores: A model that predicts football match outcomes from all over the world. *Machine Learning*, *108*, 49–75.

Dailymail (2015). Global sports gambling worth 'up to $3 trillion' | Daily Mail Online. http://www.dailymail.co.uk/wires/afp/article-3040540/Global-sports-gambling-worth-3-trillion.html. (Accessed 20 June 2018).

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *46*(2), 265–280.

Egidi, L., Pauli, F., & Torelli, N. (2018). Combining historical data and bookmakers' odds in modelling football scores. *Statistical Modelling*, *18*(5–6), 436–459.

FIFA (2017). FIFA big count 2006: 270 million people active in football - fifa.com. https://www.fifa.com/media/news/y=2007/m=5/news=fifa-big-count-2006-270-million-people-active-football-529882.html. (Accessed 23 October 2018).

Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, *21*(3), 551–564.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, *21*(2), 331–340.

Gomes, J., Portela, F., & Santos, M. F. (2016). Pervasive decision support to predict football corners and goals by means of data mining. In *New advances in information systems and technologies* (pp. 547–556). Springer.

Hubáček, O., Šourek, G., & Železnỳ, F. (2018). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, 1–19.

Hubáček, O., Šourek, G., & Železnỳ, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783–796.

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

Hughes, M., & Franks, I. (2015). *Essentials of performance analysis in sport*. Routledge.

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

Igiri, C. P. (2015). Support vector machine–based prediction system for a football match result. *IOSR Journals (IOSR Journal of Computer Engineering)*, *1*(17), 21–26.

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *52*(3), 381–393.

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

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 167–186.

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

Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, *26*(3), 471–481.

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

Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. B. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems*, *21*(7), 551–562.

Moroney, M. J. (1962). *Facts from figures*. Penguin books.

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

Owen, A. (2017). The application of hurdle models to accurately model 0-0 draws in predictive models of football match outcomes. In *Proceedings of mathsport international 2017 conference* (pp. 295).

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

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

Ryan, D., et al. (2013). *High performance discovery in time series: Techniques and case studies*. Springer Science & Business Media.

Shin, H. S. (1993). Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal*, *103*(420), 1141–1153.

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, *28*(1), 55–72.

Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International Journal of Forecasting*, *30*(4), 934–943.

Titman, A., Costain, D., Ridall, P., & Gregory, K. (2015). Joint modelling of goals and bookings in association football. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(3), 659–683.

Tüfekci, P. (2016). Prediction of football match results in Turkish Super League Games. In *Proceedings of the second international afro-european conference for industrial advancement AECIA 2015* (pp. 515–526). Springer.