Machine Learning SI manuscript No.
(will be inserted by the editor)

The Open International Soccer Database for Machine Learning

Werner Dubitzky · Philippe Lopes · Jesse Davis · Daniel Berrar

First draft: / Revision:

Abstract How well can machine learning predict the outcome of a soccer game, given the most commonly and freely available match data? To help answer this question and to facilitate machine learning research in soccer, we developed the Open International Soccer Database. Version v1.0 of the Database contains essential information from 216 743 league soccer matches from 52 leagues in 35 countries. The earliest entries in the Database are from the year 2000 when football leagues generally adopted the "three points for a win" rule. To demonstrate the use of the Database for machine learning research, we organized the 2017 Soccer Prediction Challenge. One of the goals of the Challenge was to estimate where the limits of predictability lie, given the type of match data contained in the Database. Another goal of the Challenge was to pose a real-world machine learning problem with a fixed time line and a genuine prediction task: to develop a predictive model from the Database and then to predict the outcome of the 206 *future* soccer matches taking place

W. Dubitzky
Scientific Computing Research Unit
German Research Center for Environmental Health
Helmholtz Center Munich, Germany
E-mail: werner.dubitzky@helmholtz-muenchen.de

P. Lopes
Sport and Exercise Science Department
University of Evry-Val d'Essonne, and INSERM, Paris Descartes University, France
E-mail: philippe.lopes@univ-evry.fr

J. Davis
Department of Computer Science
KU Leuven, Belgium
E-mail: jesse.davis@kuleuven.be

D. Berrar
Data Science Lab
Department of Information and Communications Engineering
Tokyo Institute of Technology, Japan
E-mail: daniel.berrar@ict.e.titech.ac.jp

from 31 March 2017 to the end of the regular season. The Open International Soccer Database is released as an open science project, providing a valuable resource for soccer analysts and a unique benchmark for advanced machine learning methods. Here, we describe the Database and the 2017 Soccer Prediction Challenge and its results.

Keywords Open International Soccer Database; 2017 Soccer Prediction Challenge; open science; soccer analytics

1 Introduction

Predicting the outcomes of sporting events has been the subject of intensive research for many years (Büchner et al., 1997). One obvious motivation for this is betting. Sports betting has become a global multi-billion-dollar industry (Forrest et al., 2005). At least since the late 1960s, statistical forecasting models have been developed for association football (Hill, 1974; Maher, 1982; Dixon and Coles, 1997; Goddard, 2005; Angelini and De Angelis, 2017), also known as *soccer*. One of the earliest studies on soccer analysis concluded that chance dominates the game (Reep and Benjamin, 1968), which makes outcome prediction very difficult.

Despite the relatively simple rules and objectives governing soccer, predicting the outcome of a soccer game is difficult. One aspect that makes soccer so popular (and unpredictable) is that goals are relatively rare and the margin of victory for the winning team is relatively low for most matches (Figure 5). Another reason why predicting the outcome of a soccer game is difficult is that goals and other game-changing circumstances (e.g., red cards, injuries, penalties) often do not occur as a result of superior or inferior play by one team, but are due to difficult-to-capture events, such as poor refereeing, unfortunate deflections or bounces of the ball, weather or ground conditions, or fraudulent match manipulation. Also, factors like political upheaval in the club's management, behavior of spectators, media pressure, and fluctuation of club player squads can influence the outcome of matches, but such events are rarely captured in databases.

To date, relatively few studies have investigated machine learning methods for soccer outcome prediction. We speculate that one reason is the lack of readily available open soccer data. Here, we present the *Open International Soccer Database* to bridge this gap.

The Database contains the most commonly and freely available as well as consistently reported information about the outcome of a league soccer match. This information concerns the goals scored by each team, teams involved, league, season, and date on which the match was played. While goals are arguably the most important match events, the drawback of such basic data is that it lacks more "sophisticated" outcome-relevant information, such as fouls committed, yellow and red cards, corners conceded by each team, or data about players, teams and clubs. Note, however, that legislation, such as the UK Data Protection Bill and the General Data Protection Regulation by the European Union, puts legal constraints to the disclosure of full names of players or coaches in publicly available databases. In sports, biometric or health data could be highly sensitive to the individual player because it can be linked to physical performance, and there are many potential misuses of such personal data, including damages to a player's reputation.

In contrast to more sophisticated data, the beauty of simple match data is that it can be easily understood and analyzed by any machine learning researcher, just like the famous Iris data set. But although the data is simple to understand, it does not mean that the scope of possible analyses is limited—on the contrary, as the special issue *Machine Learning for Soccer* shows, the data set provides considerable analytical challenges. Researchers are welcome to freely use the Database to develop and test their own strategies, methods, and tools.

However, the major motivation for developing the Open International Soccer Database was not to provide yet-another benchmark data set for the machine learning community, but to build a knowledge base that can be used for the prediction of real-world soccer matches. We therefore deliberately collected and integrated only data that are readily available for most soccer leagues worldwide, including lower leagues. In other words, this type of data is widely available, now and in the future. This is a very important aspect. If, on the other hand, the Database included highly specialized data, then its usability would be severely hampered. For example, in order to make predictions for next weekend's games, we need to obtain an update of the latest results of teams in the league of interest, retrain our model, and then predict the future games. So typically, one has less than one week for building a predictive model. Obtaining highly sophisticated match data, specifically in such a short time frame, is not feasible. This means that a lot of sophisticated data (which does not exist for all teams) that are not updated and made public in this short time interval are useless for real-world predictions. By contrast, the type of data in the Open International Soccer Database are simple, but they are the *only* data that are widely and freely available in time to make predictions.

Keeping its future usability in mind, we designed the database to provide a very large set of precisely the data that are widely and publicly available on a regular basis for practically all soccer leagues around the world. Specifically, for users who are interested in making predictions for their favorite team(s), it is important to have long historical records, as this is a key factor for reliable predictions. Simply put, such users need to add the latest results of their target teams to the Database, train predictive models, and make the desired predictions. This is of course only feasible if the historical data have the same format as the new data that the users can easily access.

The goal of the 2017 Soccer Prediction Challenge was to explore the limits of predictability using this type of widely available soccer data. More precisely, the research question could be phrased as follows: "What data on soccer matches are widely available now and in the future, and to what extent is it useful to predict match outcomes?" To address this question, we used version v1.0 of the Database as the *learning set* in the 2017 Soccer Prediction

Challenge. In January 2017, we invited the machine learning community to develop predictive models from the learning set and predict the outcome of 206 *future* matches taking place between 31 March and 9 April 2017 (Berrar et al., 2017a). In addition to exploring the limits of predictability, our intention was to pose a real "acid test" prediction challenge by requiring all participants to make their predictions *before* the actual outcome was known.

Here, we describe the Open International Soccer Database, as well as the 2017 Soccer Prediction Challenge and its results. All materials related to the Database and Challenge are publicly available under the CC0 1.0 Universal license through the Open Science Framework project sites.[1] An updated version of the Database with more entries and some corrections has already been made available at the project website. Future updates will also provide references and links to machine learning research that uses it.

This article is organized as follows. First, we review related work on soccer outcome prediction and available soccer databases. We also briefly discuss the need for reproducible research in machine learning, which motivated us to choose the Open Science Framework (OSF) (Foster and Deardorff, 2017) as accompanying repository for the Open International Soccer Database and the 2017 Soccer Prediction Challenge. Then, we describe the Database and the Challenge, and finally conclude the paper with a discussion and outlook to future work.

2 Related work

This paper is related to outcome prediction of soccer matches, sports prediction challenges, and open science. We will now position it with respect to the state of the art in these areas.

2.1 Predicting soccer match outcomes

One way to predict match results is through the use of statistical models or machine learning. Karlis and Ntzoufras (2003) used a bivariate Poisson model to predict the number of goals scored by each team in a match. Baio and Blangiardo (2010) proposed a Bayesian hierarchical model to predict the outcomes of the Italian Serie A league in the 2007/08 season. Van Haaren and Van den Broeck (2011) used kernel-based relational learning to predict the goal differences in soccer matches. Rue and Salvesen (2000) used a Bayesian approach to model the relative strength of attack and defense of a team. O'Donoghue et al. (2004) used a variety of statistical and machine learning methods to predict the results of the 2002 FIFA World Cup, but overall, only with limited success.

[1] Open International Soccer Database (Dubitzky et al., 2017), available at `https://osf.io/kqcye/`, and the 2017 Soccer Prediction Challenge (Berrar et al., 2017a), available at `https://osf.io/ftuva/`.

Another approach to predicting soccer outcomes is based on *rating systems*. Perhaps the best-known approach is an adaptation of the Elo rating system for chess (Elo, 1978), originally proposed by Arphad Elo and later adapted to soccer (Hvattum and Arntzen, 2010). The principle behind Elo-type rating schemes is that the current competitive strength of a team (or chess player) is represented by a random variable sampled from a normal or logistic probability distribution centred on the team's true strength. Comparing the distributions of two teams allows the computation of the probability of victory. The more separate the distributions, the higher the win probability of the team with the higher average rating. On the other hand, the more the distributions overlap, the greater the probability of the lower-placed team winning. The actual probability of victory is computed from the cumulative distribution function (CDF) of the difference of the teams' *rating probability density distribution*. Typically, the rating difference distribution and associated CDF are scaled somewhat arbitrarily, so that the difference of two hundred rating points equates to the higher-ranked team having a win probability of 0.75. After a match, the ratings of both teams are updated based on the actual outcome of the match and the predicted win probability.

A *probabilistic intelligence* system or pi-rating (Constantinou and Fenton, 2013) has also been used to predict soccer outcomes. At a given time point, it represents a team's strength by two numbers: the expected goal difference if the teams were to play a match against an average league team, both on the home pitch and the away pitch. Like in the Elo system, ratings are updated after each match to account for the mismatch between the observed and predicted goal differences.

2.2 Sports prediction challenges

There is a richer tradition of prediction challenges in other sports. For the past several years, Kaggle has hosted a challenge to predict the winners of all the games played in the NCAA Men's College Basketball tournament.[2] This is a single-elimination tournament involving 68 teams. The goal is to predict the winner of each match prior to the start of the tournament.

Last year, there was also a prediction challenge centered around the 2017 UEFA European Championships (Euro 2016) men's soccer tournament.[3] Euro 2016 involved 24 teams and contained a group stage followed by a knock-out stage. This prediction competition featured two tasks. The first was to predict for each nation the probability of winning, losing, and drawing if that team were to play each of the 23 other countries participating in the tournament. This task was evaluated by comparing the predicted match outcomes to actual results for the matches that ended up being played in the tournament. The second task was to predict the probability that each team would reach a given

[2] 2017 edition: `https://www.kaggle.com/c/march-machine-learning-mania-2017`

[3] `https://eu16prediction.cs.kuleuven.be/`

stage (group, round of 16, quarterfinals, etc.) in the tournament. Again, these predictions had to be submitted prior to the start of the tournament.

These challenges differ from the Open International Soccer Database and associated 2017 Soccer Prediction Challenge described in this paper in several ways. First, both our Database and Challenge are based on regular league soccer only. Other soccer games and competitions, such as tournaments of national teams and clubs, friendly games, etc., are not covered. The Challenge task was to predict the outcome of the next match of the teams for leagues that met certain conditions at the Challenge deadline by the end of March 2017. Predicting the outcome of multiple matches of the same team was not part of the Challenge. The task of the Challenge was to construct models based on the Challenge learning set only, which is identical to v1.0 of the Database. Previous challenges in this area left it open to the participants what data to use.

2.3 Publicly available resources and reproducibility

One reason for the relatively low number of data-driven studies in soccer might be the lack of publicly available databases. Data on soccer results are of course available from various online sources.[4] To our knowledge, one of the most comprehensive open databases for soccer analytics is the European Soccer Database (Mathien, 2017), an SQL database of about 25 000 soccer matches from the top leagues of 11 countries, covering seasons from 2000 to 2016. In addition to match statistics (e.g., goals, ball possession, corners, cards, etc.), this database also includes data about team formations and statistics for over 10 000 players. This database is hosted at Kaggle[5] and specifically designed for machine learning analyses. Kaggle is an interesting open data platform whose mission is to bring together data, people, discussion, and code.

Although the importance of replicablility and reproducibility has been pointed out for many years (Hirsh, 2008; Drummond, 2009; Manolescu et al., 2008; Vanschoren et al., 2012; Berrar, 2017; Berrar et al., 2017b), we believe that these issues have not yet received due attention in the machine learning community. For example, the UCI Machine Learning Repository (Lichman, 2013) hosts numerous benchmark data sets, but no analytical results, experiments, reproducible code, or any other materials that establish a context, so that pertinent questions remain open, including: How have the data been pre-processed, analyzed, and perhaps enriched so far? What is the state-of-the-art performance on these data sets? Which analytical approaches did *not* work (i.e., negative results that are usually not published)? Vanschoren et al. (2012) deplored the immense effort that is required to replicate earlier studies on benchmark data sets, simply because in practice it is not feasible to publish all details about the experiments.

[4] For example, `http://www.football-data.co.uk/`

[5] `https://www.kaggle.com/hugomathien/soccer`

There are open source software repositories for machine learning, such as *Machine Learning Open Source Software* (MLOSS)[6]. Another repository is OpenML, an open platform for hosting data sets, code, and analytical workflows, with the aim to facilitate reproducible research in machine learning (Vanschoren et al., 2013). For each project, OpenML also provides visualization tools (e.g., boxplots), a wiki, user discussions, and *tasks*, which are machine-readable containers for data subsamples (training and test sets). Furthermore, OpenML is integrated with machine learning environments such as Weka (Hall et al., 2009). OpenML is a very interesting platform; however, only a limited number of data formats are currently supported (e.g., ARFF for tabular data).

In contrast to Kaggle, the Open Science Framework (OSF)[7] is maintained by a non-profit organization, the Center for Open Science (Foster and Deardorff, 2017). OSF supports reproducible research by providing a user-friendly, free archive for data, experimental protocols, supplementary materials, code, etc. and allows the generation of persistent identifiers, i.e., digital object identifiers (DOI) and archival resource keys (ARK), which make projects citable resources. Like Kaggle, OSF, too, has a discussion board. Last but not least, OSF provides a very user-friendly frontend. In our view, OSF currently offers the simplest solution to scientists who wish to host all relevant materials in one public, citable repository. We therefore decided to make the Open International Soccer Database and all materials related to the 2017 Soccer Prediction Challenge available at OSF.

3 Contents of the Open International Soccer Database

National soccer associations organize their leagues and league format in different ways. The Open International Soccer Database includes only top leagues, and the number of top leagues covered per country varies. For example, the Database covers the top five English leagues (Premier League, Championship, League One, League Two, and National League), whereas only the top three leagues (1st Bundesliga, 2nd Bundesliga, and the 3rd Liga) from Germany are covered. For many countries, only a single league (i.e., the country's top league) is included in the Database (Table 1).

We have been collecting and manually curating soccer data from various soccer websites since 2001. For clubs whose name changed over the time frame covered in the Database (e.g., due to ownership changes), we tried to keep a single, canonical club/team name so as to ensure a continuous time series of results. Version v1.0 of the Database contains a total of 1470 team or club names. We harmonized team names so that a single team name could be used for clubs whose team name has been changing over the period covered by the Database. There are no duplicate team names across leagues and countries. For example, the three "Arsenals" in the Database are distinguished as follows:

[6] `http://mloss.org/software/`

[7] `https://osf.io/`

Arsenal (England), Arsenal Sarandi (Argentina), and Arsenal Tula (Russia). To address portability issues, team names containing accents, umlauts, or other "difficult" characters were replaced by approximations using only the 26 basic letters of the English alphabet.

In total, Version v1.0 of the Database contains results of 216 743 games played between 19/03/2000 and 21/03/2017 from 52 leagues in 35 countries. The earliest entries are from the year 2000, making sure that all entries are consistent with the "3 points per victory" rule, which some countries adopted only in the late 1990s. The most recent seasons are those that started in 2017. Table 1 gives a summary of the countries, leagues, and the number of games that were used for the 2017 Soccer Prediction Challenge.

Each entry in the Database is described by the nine fields shown below:

Sea Season in which the match was played. For example, the field value *16-17* refers to a season that started in 2016 and ended either in 2016 or 2017, depending on the league.

Lge Country and league in which the game was played. For example, the value *ENG2* refers to the second division or league in the English league system.

$Date$ Date on which the game was played.

HT Name of the *home team*.

AT Name of the *away team*.

HS Number of goals scored by *home team*.

AS Number of goals scored by *away team*.

GD Goal difference, defined as $GD = HS - AS$.

WDL Outcome of the game in terms home win (W), draw (D), and away win or loss (L).

Table 2 shows an excerpt of the Database. The last game describes the English Premier League match between Manchester City and Liverpool in the 2016/17 season. The match was played on 19/03/2017 and it ended in a 1:1 draw; hence, the goal difference, GD, is zero and the outcome, WDL, is a draw, D.

The basic format of regular league soccer is a *season*. The most common season format is double round robin where each team in a league plays every other team twice, once at each team's home venue. After the end of the season, there may be relegation, promotion or championship play-off games. In a split-season format, midterm champions may be crowned halfway through the season. Thus, the 216 743 matches in the Database are organized as consecutive season/league blocks, each block describing the date-ordered matches within a league and season. Most season/league blocks in the Database are *completed* seasons, i.e., all the matches played within one league and season.

Table 1 Summary of countries, divisions (leagues), and number of games per country covered in the Database.

Country	Code	Lge1	Lge2	Lge3	Lge4	Lge5	N
Algeria	DZA	DZA1	-	-	-	-	2675
Argentina	ARG	ARG1	-	-	-	-	6245
Australia	AUS	AUS1	-	-	-	-	2104
Austria	AUT	AUT1	-	-	-	-	3010
Belgium	BEL	BEL1	-	-	-	-	4610
Brazil	BRA	BRA1	BRA2	-	-	-	7473
Chile	CHL	CHL1	-	-	-	-	4061
China	CHN	CHN1	-	-	-	-	3694
Denmark	DNK	DNK1	-	-	-	-	3350
Ecuador	ECU	ECU1	-	-	-	-	2464
England	ENG	ENG1	ENG2	ENG3	ENG4	ENG5	40 068
Finland	FIN	FIN1	-	-	-	-	2618
France	FRA	FRA1	FRA2	FRA3	-	-	15 314
Germany	GER	GER1	GER2	GER3	-	-	13 254
Greece	GRE	GRE1	-	-	-	-	3873
Israel	ISR	ISR1	-	-	-	-	2825
Italy	ITA	ITA1	ITA2	-	-	-	13 651
Japan	JPN	JPN1	JPN2	-	-	-	6682
Mexico	MEX	MEX1	-	-	-	-	4675
Morocco	MAR	MAR1	-	-	-	-	2570
New Zealand	NZL	NZL1	-	-	-	-	722
Norway	NOR	NOR1	-	-	-	-	2102
Portugal	POR	POR1	-	-	-	-	4602
Republic of Korea	KOR	KOR1	-	-	-	-	2838
Russian Federation	RUS	RUS1	RUS2	-	-	-	6707
Scotland	SCO	SCO1	SCO2	SCO3	SCO4	-	12 712
South Africa	ZAF	ZAF1	-	-	-	-	3043
Spain	SPA	SPA1	SPA2	-	-	-	14 081
Sweden	SWE	SWE1	-	-	-	-	3616
Switzerland	CHE	CHE1	-	-	-	-	2825
The Netherlands	HOL	HOL1	-	-	-	-	5139
Tunisia	TUN	TUN1	-	-	-	-	2117
Turkey	TUR	TUR1	-	-	-	-	4250
United States of America	USA	USA1	USA2	-	-	-	3471
Venezuela	VEN	VEN1	-	-	-	-	3302
Total							216 743

Code: 3-letter country code. LgeX: Code of top X divisions or leagues in country.

N: Number of matches covered in Database.

Table 2 Excerpt of ten matches in the Database from the 2016/17 season of the English Premier League.

Sea	Lge	Date	HT	AT	HS	AS	GD	WDL
...	...	...	...	...	...	...	...	...
16-17	ENG1	18/03/2017	West Bromwich Albion	Arsenal	3	1	2	W
16-17	ENG1	18/03/2017	Crystal Palace	Watford	1	0	1	W
16-17	ENG1	18/03/2017	Everton	Hull City	4	0	4	W
16-17	ENG1	18/03/2017	Stoke City	Chelsea	1	2	-1	L
16-17	ENG1	18/03/2017	Sunderland	Burnley	0	0	0	D
16-17	ENG1	18/03/2017	West Ham United	Leicester City	2	3	-1	L
16-17	ENG1	18/03/2017	Bournemouth	Swansea City	2	0	2	W
16-17	ENG1	19/03/2017	Middlesbrough	Manchester United	1	3	-2	L
16-17	ENG1	19/03/2017	Tottenham Hotspur	Southampton	2	1	1	W
16-17	ENG1	19/03/2017	Manchester City	Liverpool	1	1	0	D
...	...	...	...	...	...	...	...	...

Sea: Season. Lge: 3-letter code for country and league/division from top.
Date: Date in DD/MM/YYYY format. HT/AT: Name of home/away team.
HS/AS: Goals scored by home/away team. GD: Goal difference as HS - AS.
WDL: Match result: W=home win, D=draw, L=away win (loss).

Season/league blocks of leagues that were in progress on 21/03/2017 show the state of that league and season on that date. Table 2 shows an example of the end of such an incomplete season/league block. Table 3 shows all 808 season/league blocks in the version v1.0 of the Database and the number of games in each block. Note that some league/season blocks are missing because of data quality issues. We will add these blocks in future versions of the Database.

The basic statistics for each league in the Database are shown in Table 4. The total proportion of outcomes across all matches in the Database is also illustrated in the boxplot of Figure 1b.

The distribution of the 25 top most frequent scorelines (out of a total of 76 distinct scores) in the Database is illustrated in the barchart of Figure 1a. The top nine scores account for 157 047 (72.46%) of all scores in the Database. The boxplots in Figure 1b shows the prior probabilities of a home win (Win), a draw (Draw), and an away win (Loss) in the Database. Clearly, a home win is by far the most common result, which is a reflection of the strong home advantage in soccer.

One aspect that makes soccer so popular (and prediction based on goals alone so difficult) is that the final outcome of the majority of soccer matches is uncertain until the end. This is because goals are relatively rare, and the margin of victory for the winning team is relatively low for most matches

Table 3 Breakdown of league/season blocks in the Database. The column names 00-01, 00-02, etc. refer to season 2000/01, 2001/02, etc.

Lge	00-01	01-02	02-03	03-04	04-05	05-06	06-07	07-08	08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	17-18	N
ARG1	380	380	380	380	378	380	380	379	380	380	380	380	380	380	450	240	238	0	6245
AUS1	210	156	156	156	0	84	84	84	84	135	165	135	135	135	135	135	115	0	2104
AUT1	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180	130	0	3010
BEL1	306	306	272	306	306	306	306	306	306	210	240	240	240	240	240	240	240	0	4610
BRA1	0	0	0	0	552	462	380	380	380	380	380	380	380	380	380	380	379	0	5193
BRA2	0	0	0	0	0	0	0	0	0	0	0	380	380	380	380	380	380	0	2280
CHE1	132	132	132	180	162	180	180	180	180	180	180	162	180	180	180	180	125	0	2825
CHL1	0	0	0	0	306	380	342	420	361	306	306	306	306	306	306	240	176	0	4061
CHN1	182	182	210	210	132	182	210	210	240	240	240	240	240	240	240	240	240	16	3694
DNK1	198	198	198	198	198	198	198	198	198	198	198	198	198	198	198	198	182	0	3350
DZA1	0	0	0	0	0	0	240	240	272	305	240	240	240	240	240	240	178	0	2675
ECU1	0	0	0	0	0	0	0	180	192	204	264	264	264	264	264	264	264	40	2464
ENG1	380	380	380	380	380	380	380	380	380	380	380	380	380	380	380	380	283	0	6363
ENG2	552	552	552	552	552	552	552	552	552	552	552	552	552	552	552	552	455	0	9287
ENG3	552	552	552	552	552	552	552	552	552	552	552	552	552	552	552	552	461	0	9293
ENG4	552	552	552	552	552	552	552	552	552	552	552	552	552	552	552	552	455	0	9287
ENG5	0	0	0	0	0	462	552	552	552	506	552	552	552	0	552	552	454	0	5838
FIN1	0	0	0	182	182	182	156	182	182	182	182	198	198	198	198	198	198	0	2618
FRA1	306	306	380	380	380	380	380	380	380	380	380	380	380	380	380	380	299	0	6231
FRA2	380	380	380	380	380	380	380	380	380	380	380	380	380	380	380	380	300	0	6380
FRA3	0	0	0	0	0	0	0	0	0	380	420	380	380	306	306	306	225	0	2703
GER1	306	306	306	306	306	306	306	306	306	306	306	306	306	306	306	306	225	0	5121
GER2	0	306	306	306	306	306	306	306	306	306	306	306	306	306	306	306	225	0	4815
GER3	0	0	0	0	0	0	0	0	380	380	380	380	380	380	380	380	278	0	3318
GRE1	240	182	240	240	240	240	240	240	240	240	240	240	240	306	305	0	200	0	3873
HOL1	306	306	306	306	306	306	306	306	306	306	306	306	306	306	306	306	243	0	5139
ISR1	0	0	0	198	198	198	198	198	198	240	240	240	182	182	182	182	189	0	2825
ITA1	306	306	306	306	380	380	380	380	380	380	380	380	380	380	380	380	290	0	6074
ITA2	380	380	380	552	462	462	462	462	462	462	462	462	462	462	462	462	341	0	7577
JPN1	0	0	0	0	240	306	306	306	306	306	306	306	306	306	306	306	306	36	3948
JPN2	0	0	0	0	0	0	0	0	0	0	0	380	462	462	462	462	462	44	2734
KOR1	0	0	0	0	156	156	182	182	182	210	210	240	352	266	228	228	228	18	2838
MAR1	0	0	0	0	0	0	240	240	240	240	240	240	240	240	240	240	170	0	2570
MEX1	0	0	380	380	306	306	306	306	306	306	306	306	306	306	306	306	243	0	4675
NOR1	0	0	0	0	0	0	0	0	182	240	240	240	240	240	240	240	240	0	2102
NZL1	0	0	0	0	0	0	84	84	56	56	56	56	56	56	72	56	90	0	722
POR1	306	306	306	306	306	306	240	240	240	240	240	240	240	240	306	306	234	0	4602
RUS1	240	240	240	240	240	240	240	240	240	240	240	240	240	240	240	240	160	0	4000
RUS2	0	0	0	0	0	0	0	0	0	380	380	379	272	342	306	380	268	0	2707
SCO1	228	228	228	228	228	228	228	228	228	228	228	228	228	228	228	228	174	0	3822
SCO2	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180	142	0	3022
SCO3	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180	143	0	3023
SCO4	180	180	180	180	180	180	180	180	180	180	180	0	180	180	180	180	145	0	2845
SPA1	380	380	380	380	380	380	380	380	380	380	380	380	380	380	380	380	279	0	6359
SPA2	462	462	462	462	462	462	462	462	462	462	462	462	462	462	462	462	330	0	7722
SWE1	182	182	182	182	182	182	182	182	240	240	240	240	240	240	240	240	240	0	3616
TUN1	0	0	0	0	0	0	182	182	182	181	181	240	112	240	240	240	137	0	2117
TUR1	306	306	306	306	306	306	306	0	306	272	306	306	306	0	306	306	0	0	4250
USA1	0	0	0	0	0	0	0	0	210	225	240	306	323	323	323	340	340	32	2662
USA2	0	0	0	0	0	0	0	0	0	0	0	112	112	98	135	165	187	0	809
VEN1	0	0	0	0	0	0	0	306	306	306	305	306	306	306	306	380	410	65	3302
ZAF1	0	0	0	0	240	240	240	240	240	240	240	240	240	240	240	240	163	0	3043
																			216 743

Lge: 3-letter code for country and league/division from top. nn-mm: Season code. Number of matches of league covered in Database.

Table 4 Basic statistics of leagues covered in the Database.

Lge	W%	D%	L%	HSg	ASg	N	Lge	W%	D%	L%	HSg	ASg	N
ARG1	43.65	29.64	26.71	1.37	1.04	6245	ISR1	41.95	27.68	30.37	1.35	1.10	2825
AUS1	45.77	24.43	29.80	1.60	1.25	2104	ITA1	46.21	27.64	26.14	1.50	1.11	6074
AUT1	47.61	25.71	26.68	1.64	1.14	3010	ITA2	44.69	32.08	23.23	1.39	1.01	7577
BEL1	47.72	24.60	27.68	1.63	1.19	4610	JPN1	42.53	24.14	33.33	1.49	1.27	3948
BRA1	50.80	25.80	23.40	1.60	1.05	5193	JPN2	39.58	27.69	32.74	1.27	1.13	2734
BRA2	50.18	25.22	24.61	1.51	1.04	2280	KOR1	40.87	28.96	30.16	1.35	1.15	2838
CHE1	46.87	24.60	28.53	1.73	1.27	2825	MAR1	43.04	35.95	21.01	1.14	0.78	2570
CHL1	47.16	24.33	28.52	1.67	1.26	4061	MEX1	44.96	28.43	26.61	1.56	1.17	4675
CHN1	47.13	29.40	23.47	1.50	1.03	3694	NOR1	48.57	24.83	26.59	1.72	1.22	2102
DNK1	43.25	25.85	30.90	1.54	1.26	3350	NZL1	46.95	15.10	37.95	1.90	1.69	722
DZA1	54.36	29.16	16.49	1.36	0.75	2675	POR1	46.02	26.25	27.73	1.42	1.06	4602
ECU1	48.13	26.54	25.32	1.47	0.99	2464	RUS1	45.70	27.43	26.88	1.40	1.03	4000
ENG1	46.46	25.65	27.90	1.53	1.13	6363	RUS2	45.59	27.11	27.30	1.30	0.98	2707
ENG2	43.91	27.65	28.44	1.46	1.12	9287	SCO1	43.43	23.70	32.86	1.48	1.19	3822
ENG3	43.91	27.11	28.98	1.48	1.15	9293	SCO2	41.89	26.74	31.37	1.47	1.23	3022
ENG4	42.70	27.63	29.67	1.42	1.12	9287	SCO3	42.64	22.96	34.40	1.58	1.38	3023
ENG5	43.53	25.64	30.83	1.50	1.20	5838	SCO4	43.55	21.48	34.97	1.57	1.35	2845
FIN1	44.50	25.82	29.68	1.49	1.15	2618	SPA1	48.09	24.61	27.30	1.57	1.13	6359
FRA1	46.25	28.79	24.96	1.40	0.97	6231	SPA2	44.37	30.15	25.49	1.37	1.00	7722
FRA2	45.27	31.96	22.77	1.36	0.93	6380	SWE1	44.80	25.86	29.34	1.54	1.18	3616
FRA3	43.32	30.15	26.53	1.37	1.02	2703	TUN1	46.62	30.04	23.33	1.26	0.85	2117
GER1	46.83	24.47	28.71	1.64	1.22	5121	TUR1	46.68	25.27	28.05	1.57	1.20	4250
GER2	45.75	27.12	27.12	1.55	1.16	4815	USA1	49.47	27.27	23.25	1.59	1.09	2662
GER3	44.64	28.36	27.00	1.46	1.07	3318	USA2	46.60	28.55	24.85	1.53	1.10	809
GRE1	49.94	24.97	25.10	1.45	0.96	3873	VEN1	47.21	28.07	24.71	1.47	1.04	3302
HOL1	47.93	23.43	28.64	1.76	1.27	5139	ZAF1	40.55	30.82	28.62	1.27	1.03	3043
Totals (sums and averages):								**45.42**	**27.11**	**27.47**	**1.48**	**1.11**	**216 743**

Lge: Name of league. Percentage proportion of home wins (W%), draws (D%) and away wins (L%).
HSg: Average goals scored by home teams. ASg: Average goals scored by away teams.
N: Number of matches in the Database.

(Figure 5). From the Database, we estimate the average number of home and away goals as 1.48 and 1.11, respectively (see Table 4). This means that, on average, the home team prevails over its opponent by a margin of 0.372 goals, reflecting the home advantage in league soccer. Moreover, when we look at the distribution of the margin of victory in the Database, we find that 86.71% of all matches end either in a draw or a victory of either team by a margin of 2 or fewer goals difference, and 95.47% are either draws or a win by either teams of 3 or fewer goals (Figure 5a).

Note that the Database captures actual outcomes of regular soccer league matches. Decisions by soccer associations that changed the result *after* the actual match are not captured in the Database. The Database may contain errors—these will be corrected once they have become known and reflected in future versions.

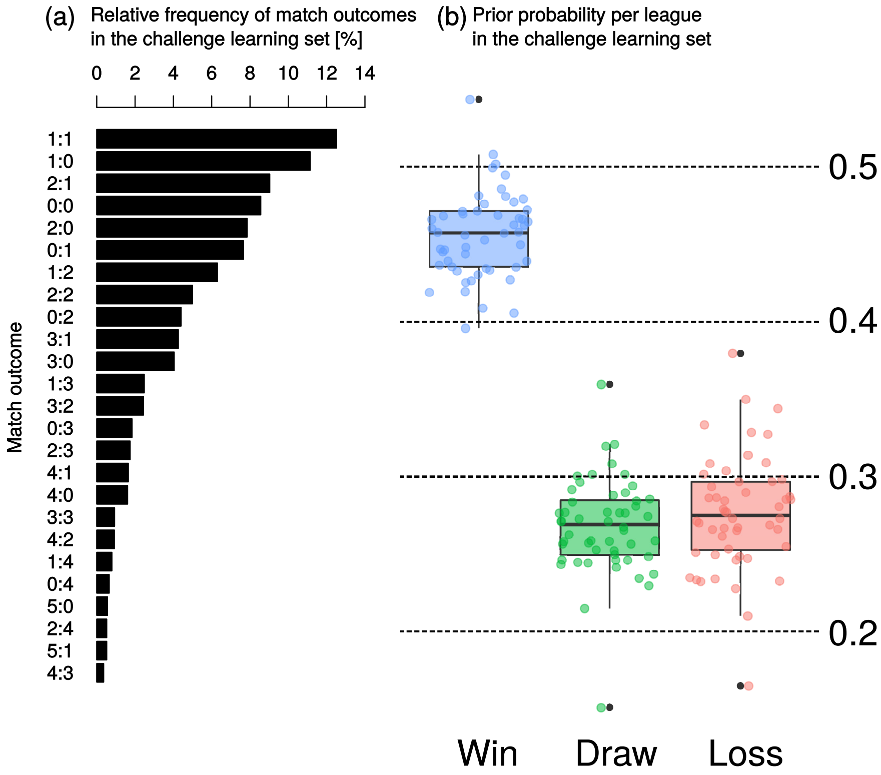

Fig. 1 (a) The 25 most frequent match scorelines in the Open International Soccer Database. The least frequent outcomes, which were observed only once in all 216 743 matches, in the Database (not shown in the barchart) are 9:4, 5:7, 11:3, and 0:10. (b) Boxplots of the prior probabilities of home team win, drawing, and losing in the Database.

4 The 2017 Soccer Prediction Challenge

The Open International Soccer Database comprises 216 743 entries, each describing the most commonly available and consistently reported data about the outcome of a league soccer match in terms of goals scored by each team, teams involved, and league, season and date on which the match was played. The beauty of this type of soccer data is that it is readily available for most soccer leagues worldwide, including lower leagues. Thus, an important research question is to determine the *limits of predictability for this type of data.* In order to find this out, we invited the machine learning community to develop predictive models based on the version v1.0 of the Database.

The 2017 Soccer Prediction Challenge was part of the special issue on *Machine Learning for Soccer* in the *Machine Learning* journal. The Challenge description was published together with the call for papers for this special issue

Table 5 Distribution of (a) margin of victory, (b) goals scored by home team, and (c) goals scored by away team in league soccer based on the Challenge learning set.

(a) Margin of victory

Margin	Frq	Frq [%]
0	58 760	27.11
1	84 478	38.98
2	44 689	20.62
3	18 992	8.76
4	6771	3.12
5	2142	0.99
6	667	0.31
7	184	0.08
8	43	0.02
9	13	0.01
10	4	0.00
-	-	-
	216 743	100.00

(b) Goals scored home team

HS	Frq	Frq [%]
0	50 612	23.35
1	72 675	33.53
2	52 722	24.32
3	26 013	12.00
4	10 108	4.66
5	3239	1.49
6	1020	0.47
7	270	0.12
8	60	0.03
9	18	0.01
10	5	0.00
11	1	0.00
	216 743	100.00

(c) Goals scored away team

AS	Frq	Frq [%]
0	73 760	34.03
1	77 737	35.87
2	42 107	19.43
3	16 223	7.48
4	5098	2.35
5	1339	0.62
6	383	0.18
7	71	0.03
8	17	0.01
9	7	0.00
10	1	0.00
-	-	-
	216 743	100.00

Margin: Winning goal difference; 0 means draw. HS/AS: Goals scored by home/away team.
Frq: Absolute frequency. Frq [%]: Frequency percentage.

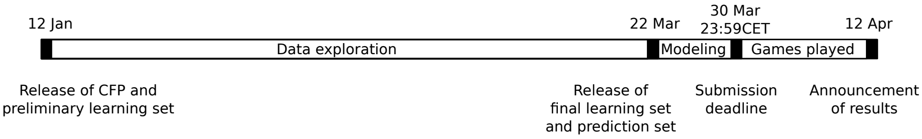

Fig. 2 Timeframe of the 2017 Soccer Prediction Challenge.

on 17 January 2017 (see supplementary material on the Challenge website[8]). Figure 2 shows the overall time frame of the challenge. The participants contacted us by email to express their interest in the Challenge and then received a web link to download the data.

Commonly in data mining competitions, the results of the prediction set are known to the competition organizers. We wanted to create a real-world prediction challenge where the outcomes referred to *real, future events* that, at the submission deadline, could not be known by *anyone*. To organize such a "real" prediction problem, we structured the Challenge around two key dates: 22/03/2017 and 30/03/2017 (Figure 2).[9] The final version of the Challenge *learning set*[10] and the *prediction set* were available on 22/03/2017. The learning set consists of data from 216 743 matches played on or *before* 22/03/2017,

[8] `https://osf.io/ftuva/`

[9] All dates and times refer to Central European Time (CET) and are expressed in the format DD/MM/YYYY.

[10] For clarity, we use here the term *learning* set instead of *training set* because the learning set may be split into training sets and validation sets by the Challenge participants.

and the 206 prediction set matches were played *after* 30/03/2017. The participants' task was to produce their final model in the time window between the two dates and submit their predictions for the prediction set by midnight CET on 30/03/2017. The particular time frame and deadline were chosen because in many leagues, regular play was suspended due to the World Cup 2018 qualifier games. Thus, there was a time window of about one week in which participants could develop their final models and apply them to the prediction set. Some lower leagues in various countries did not suspend play during this period, thus, no games from these leagues were used in the prediction set.

The final Challenge learning set is identical to v1.0 of the Open International Soccer Database presented in this article. In the remainder of this text we will use the term (final) Challenge learning set instead of Database.

4.1 Challenge data sets

We released the Challenge learning set in two instalments to the participants. The first instalment comprised data of 205 182 games (most recent entries were matches played 20/11/2016) and was released together with the public announcement of the Challenge. The participants could use the initial version of the learning set to gain an understanding of the data, try out various models, see what works and what doesn't, etc. Then, on the 22nd of March 2017, eight days before the submission deadline, the participants received the updated, final version of the learning set, together with the prediction set. The final learning set contains the results of 216 743 matches (most recent entries are matches played on 22/03/2017); it is identical to the Database presented above. It was necessary to update the learning set from its initial version so that the participants would have the match play time series of each team in the prediction set right up to the last match before their match in the prediction set.

The prediction set covers two seasons, the 2016/17 and the 2017/18 seasons, because some of the leagues covered started in 2016 and others in 2017. In total, there are 206 games for which the participants were asked to make a prediction. The prediction set has the same fields as the learning set, plus additional "x-fields." The meaning of these fields is as follows:

1. xW, xD, xL: Predicted home win, draw, and away win (loss), expressed as a real number from the unit interval $[0, 1]$, such that $xW + xD + xL = 1$. These are the fields that refer to the mandatory task of the prediction Challenge. For example, the prediction $xW = 0.7$, and $xD = 0.2$ and $xL = 0.1$ means that the model "thinks" that the probability of a home win is 0.7, the probability of a draw is 0.2, and the probability of an away win is 0.1.
2. xHS, xAS: Predicted goals scored by the home and away team, respectively, expressed as a non-negative real number. This was an optional task of the Challenge, which did not count towards the ranking of the submitted predictions.

Table 6 Excerpt of the prediction set showing the first ten matches (grouped by league). The highlighted fields xW, xD and xL had to be predicted by the Challenge participants.

Sea	Lge	Date	HT	AT	HS	AS	GD	WDL	xID	xW	xD	xL	xHS	xAS	xGD
Run	AUT1	01/04/2017	St Polten	Rapid Wien	-1	-1	0	D	1	-1	-1	-1	-1	-1	-1
Run	AUT1	01/04/2017	Austria Wien	Admira Wacker	-1	-1	0	D	2	-1	-1	-1	-1	-1	-1
Run	AUT1	01/04/2017	Sturm Graz	Wolfsberger AC	-1	-1	0	D	3	-1	-1	-1	-1	-1	-1
Run	AUT1	01/04/2017	SV Mattersburg	SV Ried	-1	-1	0	D	4	-1	-1	-1	-1	-1	-1
Run	AUT1	02/04/2017	SCR Altach	RB Salzburg	-1	-1	0	D	5	-1	-1	-1	-1	-1	-1
Run	BEL1	31/03/2017	Waregem	Anderlecht	-1	-1	0	D	6	-1	-1	-1	-1	-1	-1
Run	BEL1	01/04/2017	Charleroi	Oostende	-1	-1	0	D	7	-1	-1	-1	-1	-1	-1
Run	BEL1	02/04/2017	Gent	Club Brugge	-1	-1	0	D	8	-1	-1	-1	-1	-1	-1
Run	CHE1	01/04/2017	FC Sion	FC Thun	-1	-1	0	D	9	-1	-1	-1	-1	-1	-1
Run	CHE1	01/04/2017	FC St Gallen	FC Basel	-1	-1	0	D	10	-1	-1	-1	-1	-1	-1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Sea: Season. Lge: 3-letter code for country and league/division from top. Date: Date in DD/MM/YYYY format.
HT/AT: Name of home/away team. HS/AS: Actual goals scored by home/away team. GD: Goal difference as HS - AS.
xW/xD/xL: Predicted outcome probabilities for home win, draw, away win (loss).
xHS/xAS: Predicted goals scored by home/away team. xGD: Predicted goal difference.

3. xGD: Predicted goal difference expressed as a real number. This was another optional task of the Challenge. Note that the goal difference may not necessarily be the difference between xHS and xAS because a model might compute the goal difference without explicitly calculating the actual goals scored by each team.
4. xID: A unique identifier of the match in the prediction set.

The field *Sea* in the prediction set is *Run* for all matches, indicating that the season is in progress (referring to either the 2016/2017 or 2017/2018 season) at the time when the match entry is made. Table 6 illustrates the structure of the prediction set as it was provided to the Challenge participants. The table shows the first ten matches in the prediction, with the mandatory prediction columns are highlighted. The default for the unknown values of HS, AS, xW, xD, xL, xHS, xAS, and xGD was chosen arbitrarily and set to -1.

In order to facilitate a realistic and hard prediction challenge, the matches in the prediction set had to be carefully selected. First and foremost, we required that at the time the submissions were due (midnight, 30/03/2017 CET), the actual outcomes could not be known to *anyone* (including us, the organizers). Thus, only matches played *after* the submission deadline could be used for the prediction set. Second, since several of the leagues appearing in the learning set were not in progress[11] at the submission deadline, we could not include games from these leagues. Third, as explained in Section 1, matches from leagues that did not suspend regular league play in the period from

[11] At the Challenge deadline, the previous season in these leagues were already completed and the new season had not started yet. For example, the 2016/17 season of the FIN1 league finished on 23/10/2016 but the 2017/18 season did not start until 05/04/2017.

Table 7 Summary statistics of the prediction set matches grouped by league. Shown are the number of matches per league, the proportions of outcomes, and the averages of goals. N: Number of games in the leagues covered in the prediction set. $W\%$, $D\%$ and $L\%$: Proportion (percentage) of home wins, draws, and away wins, respectively. The actual values in the prediction set are: $W\% = 45.15$, $D\% = 26.21$ and $L\% = 28.64$. HSg and ASg: Average goals scored by home and away team, respectively.

Lge	N	W%	D%	L%	HSg	ASg	Lge	N	W%	D%	L%	HSg	ASg
AUT1	5	40.00	20.00	40.00	1.40	1.80	ISR1	7	57.14	14.29	28.57	1.00	0.86
BEL1	3	33.33	33.33	33.33	1.33	1.33	ITA1	10	20.00	30.00	50.00	1.10	1.80
CHE1	5	40.00	20.00	40.00	1.00	1.60	JPN1	9	33.33	11.11	55.56	1.44	1.67
CHN1	8	75.00	0.00	25.00	1.50	0.88	KOR1	6	33.33	50.00	16.67	1.50	1.50
ECU1	6	50.00	33.33	16.67	1.50	0.50	MEX1	9	44.44	33.33	22.22	1.11	0.89
ENG1	10	40.00	40.00	20.00	1.10	0.80	POR1	9	22.22	22.22	55.56	1.11	1.44
ENG2	12	58.33	25.00	16.67	1.25	0.58	RUS1	8	37.50	25.00	37.50	1.38	1.25
FRA1	8	50.00	25.00	25.00	1.00	0.75	SCO1	6	33.33	33.33	33.33	0.83	2.50
FRA2	10	50.00	20.00	30.00	1.50	0.90	SPA1	10	40.00	20.00	40.00	1.60	1.40
GER1	9	44.44	33.33	22.22	2.33	1.56	TUN1	3	33.33	33.33	33.33	2.00	0.33
GER2	9	33.33	33.33	33.33	1.00	1.00	USA1	10	60.00	40.00	0.00	2.10	1.00
GRE1	8	50.00	25.00	25.00	1.62	1.38	VEN1	9	44.44	44.44	11.11	1.33	0.89
HOL1	9	55.56	22.22	22.22	1.78	1.00	ZAF1	8	75.00	0.00	25.00	1.88	1.00
Totals (sums and averages):								**206**	**45.15**	**26.21**	**28.64**	**1.41**	**1.16**

Lge: Name of league. Percentage proportion of home wins (W%), draws (D%) and away wins (L%).
HSg: Average goals scored by home teams. ASg: Average goals scored by away teams.
N: Number of matches in the *prediction set*.

22/03/2017 to 30/03/2017 could not be included. For example, a full match day was played in the ENG3 league on 25/03/2017 and 26/03/2017. Fourth, each team in the prediction set had to appear only once; otherwise, the participants would have to predict outcomes of two or more matches involving the same team. Thus, only 28 of the 52 leagues from the learning set could be used to select a total of 206 matches for the prediction set.

Note that originally, we were planning to include 223 matches in the prediction set. However, during the period in which the prediction matches were being played, it turned out that some matches could not take place or were rescheduled. Hence, we contacted all Challenge participants and informed them that these matches had to be excluded due to unforeseeable circumstances. Also, because of rescheduling, some actual match dates were slightly changed.[12] Thus, the actual dates in which the prediction matches were played were from 31/03/2017 to 11/04/2017.

Table 7 shows the basic statistics of the actual outcomes and scores of the prediction matches.

[12] In total, 12 games were rescheduled, see the supplementary material at the Challenge website (Berrar et al., 2017a).

The prediction set with the outcomes for the 206 matches is provided as supplementary material at the Challenge website (Berrar et al., 2017a). The data reflect the actual outcome (observed) of the games.

Version v1.0 of the Open International Soccer Database and the learning set of the 2017 Soccer Prediction Challenge are identical. The prediction set is unique to the Challenge and not covered by v1.0 of the Database. However, as we will continue to add matches to the Database, the matches of the Challenge prediction set will be subsumed in future versions of the Database.

4.2 Performance evaluation

The task of the 2017 Soccer Prediction Challenge was to construct a model that predicts the outcomes of future soccer games based on data describing past games. We were interested in comparing the predicted probabilities for home win, draw, and away win (loss) with the actual outcomes. The commonly used Brier score (Brier, 1950), however, is not appropriate in this case because it measures only the deviance between predicted and actual scores. For example, suppose that the actual outcome is a win of the home team, which is encoded as the vector $(1,0,0)$. A model M_1 predicts $(0.6,0.3,0.1)$, whereas another model M_2 predicts $(0.6,0.1,0.3)$. The Brier score is the same for both models, but clearly, M_1 made the better prediction because it assigned a higher probability to *draw* than to *loss*, so the probability mass is shifted towards *win*.

To account for the intrinsic order in the three outcomes (win, draw, and loss), we used the *ranked probability score* (RPS) (Epstein, 1969; Constantinou and Fenton, 2012), which is defined in Equation (1),

$$\mathrm{RPS} = \frac{1}{r-1}\sum_{i=1}^{r-1}\left(\sum_{j=1}^{i}(p_j - a_j)\right)^2 \tag{1}$$

where r refers to the number of possible outcomes (here, $r = 3$ for home win, draw, and loss). Let $\mathbf{p} = (p_1, p_2, p_3)$ denote the vector of predicted probabilities for win (p_1), draw (p_2), and loss (p_3), with $p_1+p_2+p_3 = 1$. Let $\mathbf{a} = (a_1, a_2, a_3)$ denote the vector of the real, observed outcomes for win, draw, and loss, with $a_1 + a_2 + a_3 = 1$. For example, if the real outcome is a win for the home team, then $\mathbf{a} = (1,0,0)$. A rather good prediction would be $\mathbf{p} = (0.8, 0.15, 0.05)$. The smaller the RPS, the better the prediction.

The RPS value is always within the unit interval $[0,1]$. An RPS of 0 indicates perfect prediction, whereas an RPS of 1 expresses a completely wrong prediction. For example, assume that the actual, observed outcome of a soccer match was a win by the home team, coded as $A = (1,0,0)$. Let's further assume two predictions for that match: (1) a "crisp" draw prediction, B, encoded as $B = (0,1,0)$, and (2) a probabilistic prediction, C, with a home win trend, encoded as $C = (0.75, 0.20, 0.05)$. Then, by applying Equation (1), we obtain a ranked probability score of RPS $= 0.500$ for prediction B and

RPS $= 0.0325$ for prediction C. So, according to the RPS, the prediction C is better than B, which is also intuitively plausible. Or consider the prediction $D = (0.10, 0.80, 0.10)$, which leads to RPS $= 0.410$. This prediction is better than B but not as good as C.

The goal of the Challenge was to minimize the average over all ranked probability scores for all $n = 206$ matches in the Challenge prediction set,

$$\text{RPS}_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n}\text{RPS}_i \tag{2}$$

5 Results

Table 8 shows the results of the participating teams. The teams are ranked based on the average RPS according to Equation (2). Out of interest, we also calculated the average prediction accuracy rate as proportion of match outcomes in the prediction set that were correctly classified based on the highest of the three predicted outcome probabilities.

Usually, the organizers of data mining competitions are not allowed to submit their predictions, for obvious reasons. However, since the prediction set covered real future events that were not known to anyone, we also developed our own models and submitted our predictions (Berrar et al., 2018); these were nonetheless considered out-of-competition. Also, the prediction by team LJ was considered out-of-competition because it was submitted shortly *after* the deadline had already passed. The competition winners are Team OH (1st place), Team ACC (2nd place), and Team FK (3rd place). Details about their methods can be found in the special issue *Machine Learning for Soccer*.

We included the results of two baseline models or *null models*: *League Priors* and *Global Priors*. These models used only the prior information of home win, draw, and away win (loss), which were estimated from the entire learning set. 45.42% of all matches in the entire learning set ended in a win for the home team, 27.11% ended in a draw, and 27.47% ended in a win for the away team (Figure 1b and Table 4). Thus, to predict the matches in the prediction set, the *Global Priors* (*GP*) null model predicted the outcome of each match as follows: $P_{GP}(win) = 0.4542$, $P_{GP}(draw) = 0.2711$, and $P_{GP}(loss) = 0.2747$.

By contrast, the *League Priors* null model calculated the prior probabilities $P_{LG}(win)$, $P_{LG}(draw)$, and $P_{LG}(loss)$ for each of the 52 leagues individually and then used these priors as predictions for the outcomes of matches in the corresponding leagues (Figure 1b). For example, the 10 matches in the prediction set from the English Premier League (ENG1) where predicted using the league-specific priors $P_{ENG1}(win) = 0.4646$, $P_{ENG1}(draw) = 0.2565$, and $P_{ENG1}(loss) = 0.2790$ (Table 4).

Table 8 Summary of the results of the 2017 Soccer Prediction Challenge. Participating teams are ranked based on increasing values of the average ranked probability score, calculated from the submitted predictions for the 206 games of the prediction set. Shown is also the accuracy, i.e., the percentage of correctly predicted games. Submissions by the organizers (Team DBL) are out-of-competition and marked by *.

Rank	Team	$\mathrm{RPS}_{\mathrm{avg}}$	Accuracy	Method
1	Team DBL*	0.2054	0.5194	Berrar et al. (2018)
2	Team OH	0.2063	0.5243	Hubáček et al. (2018)
3	Team ACC	0.2083	0.5146	Constantinou (2018)
4	Team FK	0.2087	0.5388	Tsokos et al. (2018)
5	Team DBL*	0.2149	0.5049	Berrar et al. (2018)
6	Team HEM	0.2177	0.4660	
7	League Priors	0.2255	0.4515	Prior information based on leagues
8	Team EB	0.2258	0.4854	N/A
9	Global Priors	0.2261	0.4515	Global priors of win, draw, lose
10	Team LJ	0.2313	0.4126	N/A
11	Team AT	0.3981	0.3883	N/A
12	Team LHE	0.4515	0.3398	N/A
13	Team EDS	0.4515	0.3592	N/A

6 Discussion

Soccer is arguably the world's most popular team sport. It is also interesting from an analytical point of view because it presents unique challenges. Soccer typically involves a low number of goals, a low margin of victory (Figure 5), and difficult-to-capture events that often determine the final outcome of a match. On the other hand, data capturing the essential aspects of a match is readily available. However, soccer data is rarely available in a form directly usable by machine learning methods. Moreover, how soccer data from different countries and leagues or other competitions, such club or national team championships, should be combined to produce a larger data set suitable for machine learning is not immediately obvious. Thus, predictive modeling in soccer poses interesting challenges with respect to integration of domain knowledge and feature engineering.

To predict the number of goals scored and conceded, mostly statistical approaches, such as Poisson models, have been applied so far, whereas relatively few machine learning methods have been proposed (Kumar, 2013). We speculate that this is partly due to the lack of publicly available databases that allow data-driven analyses. We hope that the presented Open International Soccer Database will bridge this gap.

Our decision to host all materials at the OSF platform was motivated by the need for open science and reproducible research. We believe that OSF provides an excellent infrastructure for archiving all documents that are relevant for a replicable research project. Also, what makes OSF particularly interesting for academia is the fact that permanent document object identifiers can be assigned to projects, thereby making them citable resources. The Challenge participants were therefore encouraged (but not obliged) to deposit their code and predictions (in the format of Table 6) at the Challenge website.

We organized the 2017 Soccer Prediction Challenge to address the following research question: "How well can machine learning predict the outcome of soccer matches, given the type of readily available data about soccer matches?" In other words, given this type of readily available soccer data, and given a decent-size learning set like the one used in the Challenge, we can further ask: "What is the limit of predictability one can expect?"

Nine teams submitted valid predictions. Several other teams expressed an interest in the Challenge and requested the data, but did not submit their predictions in the end. Compared to many other data mining competitions, our Challenge could garner only a relatively low participation although it was widely advertised. There are several possible reasons for this low participation; for example, there was no monetary prize, and the Challenge was not organized in association with any well-known conference. In addition to the nine submissions that we received, we also included the results of our own two models, which were of course out-of-competition contributions.

Table 8 shows that, in terms of average RPS, the four top-ranked predictions are remarkably close, with a mean average RPS_{avg} of 0.2072. Thus, it is tempting to speculate that a performance of around $\text{RPS}_{\text{avg}} = 0.2072$ is close to the limit of predictability for this kind of data. We hope that the Database will be widely used by the machine learning community and that innovative methods will push this limit further.

Clearly, adding more data relevant to the outcome of a match (e.g., data about other match events, players, teams, etc.) might improve the predictability. However, as we discussed in the Introduction, one problem with additional data is that it may not be readily available for a wide range of countries and leagues, and therefore the number of matches may not be as high as in the learning set of the Challenge. Still, the question to what extent such additional data could improve the results presented in this paper is interesting and scope for future research.

Acknowledgements After we released the Challenge data sets, we received valuable feedback from the participants regarding the evaluation of the predicted outcomes. In particular, we wish to thank team ACC and team FK for their constructive comments. We also thank the three anonymous reviewers for their valuable comments.

References

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

Baio G, Blangiardo M (2010) Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics 37(2):253–264

Berrar D (2017) Confidence curves: an alternative to null hypothesis significance testing for the comparison of classifiers. Machine Learning 106(6):911–949

Berrar D, Lopes P, Davis J, Dubitzky W (2017a) The 2017 Soccer Prediction Challenge URL `http://doi.org/10.17605/OSF.IO/FTUVA`

Berrar D, Lopes P, Dubitzky W (2017b) Caveats and pitfalls in crowdsourcing research: the case of soccer referee bias. International Journal of Data Science and Analytics 4(2):143–151

Berrar D, Lopes P, Dubitzky W (2018) Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine Learning *(to appear)*

Brier G (1950) Verfication of forecasts expressed in terms of probability. Monthly Weather Review 78(1):1–3

Büchner AG, Dubitzky W, Schuster A, Lopes P, O'Donoghue PG, Hughes JG, Bell DA, Adamson K, White JA, Anderson JMCC, Mulvenna MD (1997) Corporate evidential decision making in performance prediction domains. In: Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, USA, UAI'97, pp 38–45

Constantinou A (2018) Dolores: a model that predicts football match outcomes from all over the world. Machine Learning DOI 10.1007/s10994-018-5703-7

Constantinou A, Fenton N (2012) Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. Journal of Quantitative Analysis in Sports 8(1):article 1

Constantinou AC, Fenton NE (2013) Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. Journal of Quantitative Analysis in Sports 9(1):37–50

Dixon M, Coles S (1997) Modelling association football scores and inefficiencies in the football betting market. Applied Statistics 46(2):265–280

Drummond C (2009) Replicability is not reproducibility: Nor is it good science. Proceedings of Evaluation Methods for Machine Learning Workshop at the 26th International Conference on Machine Learning, Montreal, Canada pp 1–6

Dubitzky W, Lopes P, Davis J, Berrar D (2017) The Open International Soccer Database URL `http://doi.org/10.17605/OSF.IO/KQCYE`

Elo AE (1978) The rating of chessplayers, past and present. Batsford London

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology 8(6):985–987

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

Foster E, Deardorff A (2017) Open Science Framework (OSF). Journal of the Medical Library Association 105(2):203–206

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

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I (2009) The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11(1):10–18

Hill I (1974) Association football and statistical inference. Applied Statistics 23(2):203–208

Hirsh H (2008) Data mining research: Current status and future opportunities. Statistical Analysis and Data Mining 1(2):104–107

Hubáček O, Šourek G, Železný F (2018) Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning DOI 10.1007/s10994-018-5704-6

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

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

Kumar G (2013) Machine learning for soccer analytics. Master's thesis, Departement Computerwetenschappen, KU Leuven, Belgium

Lichman M (2013) UCI Machine Learning Repository. URL `http://archive.ics.uci.edu/ml`

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

Manolescu I, Afanasiev L, Arion A, Dittrich J, Manegold S, Polyzotis N, Schnaitter K, Senellart P, Zoupanos S, Shasha D (2008) The repeatability experiment of SIGMOD 2008. ACM SIGMOD Record 37(1):39–45

Mathien H (2017) The European Soccer Database URL `https://www.kaggle.com/hugomathien/soccer`

O'Donoghue P, Dubitzky W, Lopes P, Berrar D, Lagan K, Hassan D, Bairner A, Darby P (2004) An evaluation of quantitative and qualitative methods of predicting the 2002 FIFA World Cup. Journal of Sports Sciences 22(6):513–514

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

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

Tsokos A, Narayanan S, Kosmidis G I Baio, Cucuringu M, Whitaker G, Kiràly F (2018) Modeling outcomes of soccer matches. Machine Learning *(to appear)*

Van Haaren J, Van den Broeck G (2011) Relational learning for football-related predictions. In: Proceedings of the 21st International Conference on Inductive Logic Programming (ILP-2011), Windsor Great Park, UK, pp 1–6

Vanschoren J, Blockeel H, Pfahringer B, Holmes G (2012) Experiment databases. Machine Learning 87(2):127–158

Vanschoren J, van Rijn JN, Bischl B, Torgo L (2013) OpenML: Networked science in machine learning. ACM SIGKDD Explorations Newsletter 15(2):49–60