Call for Papers for a Special Issue in the Springer Journal

Machine Learning

Machine Learning for Soccer

Description of Prediction Challenge

Special Issue Editors:

Werner Dubitzky[1], Daniel Berrar[2], Philippe Lopes[3], Jesse Davis[4]
[1] Freelance scientist, Germany, werner@dubitzky.com

[2] College of Engineering, Shibaura Institute of Technology, Japan, dberrar@shibaura-it.ac.jp

[3] Sport and Exercise Science Department, University of Evry-Val d'Essonne, and
INSERM UMRS 1124, Paris Descartes University, France, philippe.lopes@univ-evry.fr

[4] Department of Computer Science, KU Leuven, Belgium, jesse.davis@kuleuven.be

This special issue includes a machine learning research challenge based on a *training data set* containing the results of over 200,000 soccer matches. Challenge participants should use this data set to construct a model that predicts the outcome of future soccer matches contained in the *prediction data set*. This challenge presents a unique real-world machine learning prediction problem and it involves solving various machine learning tasks. In particular, there are two main issues to address. First, how should one derive predictive features from the data? There is no obvious way of doing this – many approaches are conceivable. Second, the data consists of matches from different soccer leagues around the world. How can these data be combined to maximize the size of the training data that can be used to construct a model? Again, there may be many ways of doing this.

In this document, we first present the data sets that are provided for this challenge. Then we describe how the predictions should be coded and how they will be evaluated.

Data description

Three data sets are provided for this challenge:

1. The *initial training set*. This set is provided immediately to all challenge participants.
2. The *updated training set*. This set will be made available on 22 March 2017 CET.
3. The *prediction set*. This set will be made available on 22 March 2017 CET.

There is also a test version of the *updated training set* and *prediction set*. These are provided immediately so that participants are able to test their data processing workflows accordingly.

The **initial training set** contains the match data of over 200,000 past soccer matches played in 52 soccer leagues starting from the 2000/2001 season. This data set contains match data of leagues whose season has already been completed. The challenge participants should use this data set to develop their predictive models.

The **updated training set** will be provided to all challenge participants on 22 March 2017 CET. This data set contains match data of leagues which are still running (i.e. are not completed yet) by the all-important 30 March 2017 deadline.

The **prediction set** contains matches of the running seasons in the updated training data set that are played immediately *after* the 30 March 2017 deadline. Predicting the outcome of these matches forms the prediction challenge of this special issue. Like the updated training set, the prediction set will be provided to all challenge participants on 22 March 2017 CET.

To guide model development and data processing workflows, we provide a `TEST` version of the updated training and prediction set together with the initial training set. We use the file name additions `_CHALLENGE` and `_TEST` to distinguish the two versions. Furthermore, all data sets are provided in Excel and tab-delimited text (ASCII) format. Table 1 describes the files, their format and the date when they will be made available to the challenge participants.

Table 1. Data files provided for this challenge.

File Name	**Format**	**Remark**	**Date Made Available**
InitialTrain_CHALLENGE.xlsx InitialTrain_CHALLENGE.txt	Excel 2007 Text	Data from over 200,000 matches from 52 leagues. Completed seasons only.	Immediately
UpdatedTrain_CHALLENGE.xlsx UpdatedTrain_CHALLENGE.txt	Excel 2007 Text	Data from leagues whose season is still running at the 30 March 2017 deadline.	on/before 22 March 2017
Predict_CHALLENGE.xlsx Predict_CHALLENGE.txt	Excel 2007 Text	Data from leagues whose season is running at the 30 March 2017 deadline.	on/before 22 March 2017
UpdatedTrain_TEST.xlsx UpdatedTrain_TEST.txt	Excel 2007 Text	Data from leagues whose season is still running at 11 December 2016.	Immediately
Predict_TEST.xlsx Predict_TEST.txt	Excel 2007 Text	Data from leagues whose season is running at 11 December 2016.	Immediately

Please contact the guest editors to get access to the data.

Initial Training Data Set

The *initial train set* contains 205,182 rows, where each row corresponds to a soccer match as illustrated in Table 2 below.

Table 2. First and last six rows in initial training data set.

```
         Sea  Lge       Date             HT             AT  HS AS GD  WDL
1      00-01 GER1 11/08/2000       Dortmund  Hansa Rostock   1  0  1    W
2      00-01 GER1 12/08/2000  Bayern Munich  Hertha Berlin   4  1  3    W
3      00-01 GER1 12/08/2000       Freiburg  VfB Stuttgart   4  0  4    W
4      00-01 GER1 12/08/2000   Hamburger SV    Munich 1860   2  2  0    D
5      00-01 GER1 12/08/2000 Kaiserslautern         Bochum   0  1 -1    L
6      00-01 GER1 12/08/2000     Leverkusen      Wolfsburg   2  0  2    W
...
205177 16-17 FIN1 23/10/2016 IFK Mariehamn   Ilves Tampere   2  1  1    W
205178 16-17 FIN1 23/10/2016     Vaasan PS        FC Lahti   0  1 -1    L
205179 16-17 FIN1 23/10/2016    Kuopion PS      Kemi Kings   1  0  1    W
205180 16-17 FIN1 23/10/2016  Rovaniemi PS Helsingfors IFK   0  0  0    D
205181 16-17 FIN1 23/10/2016   PK35 Vantaa     Inter Turku   2  1  1    W
205182 16-17 FIN1 23/10/2016  HJK Helsinki       Seinajoki   0  0  0    D
```

Each season/league block of rows describes all matches played in that season and league. The match entries within each season/league block are sorted by date (the most recent match at the bottom of the block). This is illustrated in Table 3, which shows six of the 552 matches at the beginning and end of the 2004/2005 season of the top league in Brazil (indicated as BRA1).

Table 3. Excerpt of matches of the BRA1 league in 2004/2005 season.

```
      Sea  Lge       Date                            HT                          AT HS AS GD WDL
1   04-05 BRA1 21/04/2004                   Figueirense               Internacional  1  0  1   W
2   04-05 BRA1 21/04/2004       Botafogo Rio de Janeiro         Goias Esporte Clube  1  4 -3   L
3   04-05 BRA1 21/04/2004                   Sao Caetano    Esporte Clube Vitoria BA  1  0  1   W
4   04-05 BRA1 21/04/2004        Cruzeiro Esporte Clube     Esporte Clube Juventude  2  1  1   W
5   04-05 BRA1 21/04/2004 Sociedade Esportiva Palmeiras            Atletico Mineiro  0  0  0   D
6   04-05 BRA1 22/04/2004           Gremio Porto Alegre     Flamengo Rio de Janeiro  0  0  0   D
...
547 04-05 BRA1 19/12/2004 Esporte Clube Vitoria BA                      Ponte Preta  1  2 -1   L
548 04-05 BRA1 19/12/2004  Flamengo Rio de Janeiro           Cruzeiro Esporte Clube  6  2  4   W
549 04-05 BRA1 19/12/2004      Goias Esporte Clube                     Sao Paulo FC  2  0  2   W
550 04-05 BRA1 19/12/2004   Criciuma Esporte Clube                      Coritiba FC  3  3  0   D
551 04-05 BRA1 19/12/2004 Fluminense Rio de Janeiro   Sociedade Esportiva Palmeiras  1  1  0   D
552 04-05 BRA1 19/12/2004            Internacional                     Parana Clube  2  1  1   W
```

The important thing to remember is that the initial training set contains all the fixtures played in a league over an entire season (see also notes at the end of this document).

The following field names appear in all three data sets (*initial training set*, *updated training*, and *prediction set*) and have the following meaning:

Sea: Defines the season in which the match was played using the *yy-yy* format. For example, the label *00-01* refers to the 2000/2001 season. Notice, we always use two consecutive calendar years in the season label, even when the season started and finished in the same calendar year. The first number of the season label always refers to the year in which the season started.

Lge: Defines the league name using a 3-letter code (for the country) followed by a number (identifying the order of the league from the top). For example, *ENG1* refers to the English Premier League and *ITA2* refers to the Italian Serie B league.

Date: Defines the date when the match was played using the *dd/mm/yyyy* format.

HT: Defines the name of the *home* team.

AT: Defines the name of the *away* team.

HS: Stands for *home* score and refers to the number of goals scored by the *home* team.

AS: Stands for *away* score and refers to the number of goals scored by the *away* team.

GD: Stands for goal difference and is defined as follows: *GD* = *HS* – *AS*. Thus, the goal difference is asymmetric – a positive value indicates a victory of the home team, a negative value indicates a victory of the away team, a value of zero indicates a draw.

WDL: Stands for win, draw, loss – such that *W* indicates a victory of the home team, *D* indicates a draw, and *L* indicates a victory of the away team.

It should be quite easy for analysts to “deconstruct” the season/league blocks in the data in a way that is needed for their model construction.

Updated Training Data Set

The *updated training set* is formatted identically (same fields) to the *initial training set*. The only difference is that the *updated training set* contains match info of leagues whose season has started *before* but is not yet completed *at* the 30 March 2017 deadline. This means that the season in these leagues is still running at the time the predictions are to be made – hence we use the label *Run* in the *Sea* field. Table 4 below illustrates this based on data from the English Premier League (ENG1) showing data of 12 matches of the currently running seasons (2016/2017) *before* the 12 December 2016. Of course, for the actual *updated training set* the match data will extend to just before the deadline in March 2017. The data below is taken from the *test* version of the updated training set (which contains match data up to the 11/12/2016).

Table 4. Excerpt of matches of the ENG1 league in currently running (Run) 2016/2017 season.

```
    Sea  Lge       Date                HT                   AT HS AS GD WDL
...
139 Run ENG1 04/12/2016           Everton    Manchester United  1  1  0   D
140 Run ENG1 05/12/2016     Middlesbrough            Hull City  1  0  1   W
141 Run ENG1 10/12/2016           Watford              Everton  3  2  1   W
142 Run ENG1 10/12/2016           Arsenal           Stoke City  3  1  2   W
143 Run ENG1 10/12/2016           Burnley          Bournemouth  3  2  1   W
144 Run ENG1 10/12/2016         Hull City       Crystal Palace  3  3  0   D
145 Run ENG1 10/12/2016      Swansea City           Sunderland  3  0  3   W
146 Run ENG1 10/12/2016    Leicester City      Manchester City  4  2  2   W
147 Run ENG1 11/12/2016           Chelsea West Bromwich Albion  1  0  1   W
148 Run ENG1 11/12/2016 Manchester United    Tottenham Hotspur  1  0  1   W
149 Run ENG1 11/12/2016       Southampton        Middlesbrough  1  0  1   W
150 Run ENG1 11/12/2016         Liverpool      West Ham United  2  2  0   D
```

Prediction Data Set

The *prediction set* contains additional fields that should be used by the challenge participants to register their prediction. All other fields are the same as in the *initial/updated training sets*, but the value of outcome fields (*HS*, *AS*, *GD*, *WDL*) have been set to -1 to indicate unknown values.

The meaning of the additional fields in the *prediction set* is as follows:

xID: An identifier used to label (number) the matches in the prediction set.

xW: *Predicted* win by the *home* team.

xD: *Predicted* draw.

xL: *Predicted* win by the *away* team.

xHS: *Predicted* home score (goals scored by the *home* team).

xAS: *Predicted* away score (goals scored by the *away* team).

xGD: *Predicted* goal difference.

When you receive the prediction set, the values of the *xFields* are set to -1 one to indicate unknown value, except for the *xID* field, which will be filled with an ID uniquely identifying the match. Table 5 illustrates the format of the *prediction set* by an excerpt from the *test prediction set*.

Table 5. Excerpt from test prediction set of currently (Run) running season in ENG1 league.

```
   Sea  Lge       Date                   HT                AT HS AS GD WDL xID xW xD xL xHS xAS xGD
1  Run ENG1 13/12/2016          Bournemouth    Leicester City -1 -1 -1  -1  29 -1 -1 -1  -1  -1  -1
2  Run ENG1 13/12/2016              Everton           Arsenal -1 -1 -1  -1  30 -1 -1 -1  -1  -1  -1
3  Run ENG1 14/12/2016        Middlesbrough         Liverpool -1 -1 -1  -1  31 -1 -1 -1  -1  -1  -1
4  Run ENG1 14/12/2016           Sunderland           Chelsea -1 -1 -1  -1  32 -1 -1 -1  -1  -1  -1
5  Run ENG1 14/12/2016      West Ham United           Burnley -1 -1 -1  -1  33 -1 -1 -1  -1  -1  -1
6  Run ENG1 14/12/2016       Crystal Palace Manchester United -1 -1 -1  -1  34 -1 -1 -1  -1  -1  -1
7  Run ENG1 14/12/2016 West Bromwich Albion      Swansea City -1 -1 -1  -1  35 -1 -1 -1  -1  -1  -1
8  Run ENG1 14/12/2016      Manchester City           Watford -1 -1 -1  -1  36 -1 -1 -1  -1  -1  -1
9  Run ENG1 14/12/2016           Stoke City       Southampton -1 -1 -1  -1  37 -1 -1 -1  -1  -1  -1
10 Run ENG1 14/12/2016    Tottenham Hotspur         Hull City -1 -1 -1  -1  38 -1 -1 -1  -1  -1  -1
```

Prediction Task/Evaluation Description

The *xFields* in the prediction set are prepared in such a way that 3 types of outcome predictions can be registered/recorded.

1. *Match outcome*: Predicts a win of either the home or away team, or a draw. Arguably, the match outcome prediction is the easiest of the three types of predictions.
2. *Match score*: Predicts the goals scored by the home and the away team.
3. *Match goal difference*: Predicts the goal difference of the match which is defined by the difference of the goals scored by the home team minus the goals scored by the away team.

Notice, all challenge participants MUST predict the *match outcome* prediction, as this prediction forms the basis to evaluate and rank the challenge contributions. You need to use the fields *xW*, *xD* and *xL* to code the prediction in the *prediction set* and submit the *prediction set* with the coded predictions before the deadline on 30 March 2017.

The prediction of the *match score* (fields: *xHS*, *xAS*) and *match goal difference* (field: *xGD*) is OPTIONAL. You do not need to fill in the corresponding fields in the prediction set (but you are free to do so). Clearly, if your method predicts the match score you are able to derive both the goal difference and the outcome, and if your method predicts the goal difference, you can derive the outcome.

Match Outcome Prediction

All challenge participants are required to register their match outcome prediction in the fields *xW*, *xD* and *xL* as follows:

$$xW, xD, xL \in [0,1]$$
$$xW + xD + xL = 1$$

In other words, you code your match outcome prediction by three numbers drawn from the unit interval such that the sum of the three numbers equals 1. For example, (*xW*=1, *xD*=0, *xL*=0) would code a home win if a discrete classifier has predicted a home win. Another example uses three non-zero values to predict the match outcome as follows: (*xW*=0.75, *xD*=0.20, *xL*=0.05). Notice, it is important that the three values add up to 1.00.

We will use the following codes to represent the *actual match outcomes* of the matches in the *prediction set* after the matches have been completed after the 30 March 2017 deadline:

Win by *home* team:	(*aW*=1, *aD*=0, *aL*=0)
Draw:	(*aW*=0, *aD*=1, *aL*=0)
Win by *away* team:	(*aW*=0, *aD*=0, *aL*=1)

Using this coding, we will use the ***ranked probability score* (*RPS*)** to determine the error between the actual observed outcome of a match and the prediction. For a single match the RPS is computed as follows:

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

or, in an alternative but equivalent formulation;

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

where

r is the number of outcomes ($r = 3$ in the case of soccer);

p_j is the *predicted* outcome at position j, such that $p_j \in [0,1]$ for $j = 1,2,3$ and $p_1 + p_2 + p_3 = 1$.

a_j is the *observed* outcome at position j, such that $a_j \in \{0,1\}$ for $j = 1,2,3$ and $a_1 + a_2 + a_3 = 1$. That is (1,0,0) for home win, (0,1,0) for draw, and (0,0,1) for away win.

An RPS of zero indicates a perfect prediction and an RPS of 1 a completely wrong prediction.

For example, given the observed outcome was a win by the home team, coded as (*xW*=1, *aD*=0, a*L*=0) or simply (1,0,0) (representing a home win), and the predicted outcome was (*xW*=0, *xD*=1, *xL*=0) or simply (0,1,0), then the rank probability score would be *RPS* = 0.5000.

And for a (*xW*=0.75, *xD*=0.20, *xL*=0.05) as prediction with the same observed outcome (1,0,0), we would compute *RPS* = 0.0325.

The overall match outcome prediction error/score will be computed as average of the individual match outcome prediction RPS scores from the matches in the prediction set. The team producing the lowest average match outcome prediction score will be declared the winner of the prediction challenge.

Match Score Prediction

To participate in the challenge, it is *not* mandatory to provide *match score* predictions. The match score predictions, *xHS* and *xAS*, should be coded as follows:

$$xHS, xAS \in [0, \infty]$$

The following are two examples of match score predictions

$$(xHS = 1, xAS = 3)$$

$$(xHS = 0.87, xAS = 1.51)$$

Based on the actual match score, coded as (*aHS*, *aAS*), the error, *err*, of the match score prediction, (*xHS*, *xAS*), will be determined as follows

$$err = \sqrt{\frac{(aHS - xHS)^2 + (aAS - xAS)^2}{2}}$$

Match Goal Difference Prediction

To participate in the challenge, it is *not* mandatory to provide *match goal difference* predictions. The *match goal difference* predictions, *xGD*, should be provided (coded) as follows:

$$xGD \in [-\infty, \infty]$$

Based on the actual match goal difference, coded as *aGD*, the error, *err*, of the match goal difference prediction, *xGD*, will be determined as follows

$$err = \sqrt{(aGD - xGD)^2}$$

Schedule for Prediction Challenge

Date	**Activity/Milestone**
Ongoing	Participants contact the special issue editors to express/register their interest in the prediction challenge. Challenge participants get access to the *training data set* including a description of the data and prediction challenge. The training data set consists of the soccer results (league, season, date of match, home team name, away team name, home score, away score) of over 200,000 past soccer matches (all from regular league play only, not from tournaments, friendlies, or international matches).
22 March 2017	Participants receive an *updated training set* and the *prediction data set*. The *updated training set* includes the results of approximately 4000 additional matches from league play in the ongoing season. The *prediction set* contains ca. 400 matches that will be played *after 30 March 2017* and the participants should predict the outcomes of these matches. Thus, at the time when the predictions are made and submitted to the special issue editors, the outcome of these matches is unknown. This is the ultimate test for predictive machine learning models.
Midnight CET **30 March 2017**	Strict deadline for challenge participants to submit their predictions for the future matches in the prediction data set.
15 April 2017	Challenge participants are notified about their performance on the task and how they fared compared to other participants. Authors of the top-ranked predictions are invited to submit a full manuscript on their machine approach in line with the overall schedule of the special issue (see above). Those invited to submit their manuscript are also requested to apply their model to all the matches in the updated training set.

Please also not the overall schedule of the special issue:

Date	**Activity/Milestone**
15 May 2017	Deadline for manuscript submissions
15 August 2017	Review results (round 1)
15 October 2017	Revised papers due
15 November 2017	Final review results / final manuscript selection
Beginning 2018	Publication of special issue

Notes

Various teams had their scores/points adjusted retrospectively by their football association due to irregularities. The match scores in the data sets reflect the actual scores of the matches without adjustments by football associations.

For example, in the 2016-2017 the K-League (KOR1) team Jeonbuk was punished by a 9-point deduction. Such changes are not captured in the match data; instead, the actual match results are kept in the data sets.

Some clubs have changed their club name during or between seasons (e.g. due to different club ownership). We have tried to use a single club name across the seasons for such clubs to facilitate cross-season coherence of the data.

Some leagues run a split-season schedule. Generally, we have combined all matches of a league within a season into a single season/league block without differentiating season blocks across a split season. Similarly, we normally have not included play-off matches that are often played between split-season blocks to determine season-block champions.

We have tried to remove aborted matches. So there are season/league blocks which do not show the total expected number of matches due to removal of aborted matches.

The data provided must not be used for commercial purposes.