

THE APPLICATION OF THE MACHINE LEARNING PRINCIPLES IN THE SPORTS BETTING SYSTEMS

Marek RUŽIČKA, Martin CHOVANEC

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice,
Letná 9, 042 00 Košice, E-mail: marek.ruzicka@tuke.sk, martin.chovanec@tuke.sk

ABSTRACT

Use of machine learning for football match result prediction is still a difficult task. Betting companies use predictions for betting odds estimation and profit generation. Choosing the correct data for training models is thus crucial. In this paper the data model is selected and multiple machine learning algorithms are compared. Along with data model selection, impact of number of game seasons on success rate was analyzed. Last but not least, simple three class prediction was compared to separate binary predictions of win, lose and draw. All models were tested by one whole game season from last 380 games of English football championship Premier League. The best results achieved passed 60% accuracy, which outperforms all the analyzed works.

Keywords: Betting, Machine learning, Prediction, Sport

1. INTRODUCTION

Machine learning usage in sport betting systems is still a big field for a research. Multiple papers have been published in this topic already in the past, but there is still much to improve.

Michael Purucker published one of the first papers on this topic, Neural network quarterbacking [1]. He tried to predict results of matches in National Football League (NFL). His dataset contained only last season data and he tried to predict only 14 matches. Prediction success of his model was about 61%.

In 2003, Kahn continued research based on Purucker's paper [2]. He used more variables and his prediction success increased to 75%. In this paper was used similar dataset for training and testing models. To be mentioned, draw is very uncommon in rugby, which makes prediction more accurate.

McCabe and Trevathan used more complex datasets in their paper Artificial Intelligence in Sports Prediction [3]. They used data from year 2002 and tried 4 different leagues, including NFL and Premier league. Average prediction success was 54,6% for Premier league, with best score 58,9% in a single season. In NFL the score was somewhere between 63 to 67% in average. We can observe, that predictions in long term can have greater error.

Tax and Joustra used data from 13 years of German soccer [4]. They compared multiple algorithms with the best result of 56,1%. They tried to use data about bets for predictions.

In Machine learning for Soccer analytics author used over 300 variables about the team itself, along with the player statistics, to predict the result [5]. He reached success rate of 53,4%, which he justifies by the complexity of prediction of the draw.

Web portal kickoff.ai stating to use machine learning and addresses similar problem. They use Bayesian interface to predict the percentual possibilities of each possible result of soccer matches. Analyzing the last 100 matches predicted (to 22nd March 2019), kickoff has 56% success rate. In the 100 matches, no draw was predicted.

Similar result about draws was observed in data we ob-

tained. From 7830 matches over 21 seasons, only 9 draws were predicted, of which only 2 were correct.

Analysing the results mentioned above, draw prediction seems to be the crucial part of soccer result prediction. Higher success rate in draw prediction is one of the aims of this paper.

2. OBJECTIVES

There are four main objectives analyzed concurrently.

First of all, a hypothesis that long term statistics of each team will increase prediction rate is introduced. Many of the mentioned papers used short term data only. Training the model on the known data from current season and predicting upcoming matches might seem to be a good idea from the "old data might not be valid" view. However, predicting first few weeks this way is plain guessing, as there is no data.

The second objective derives from the first. We want to determine, how many seasons is the best count in training model. We can't simply expect, that using all 21 seasons we obtained is the best option. The teams change as well as the team forms change over time. Season count thus can be very important.

Objective number three is to determine, which of the machine learning algorithms is the best for this particular problem. Five different algorithms were tried, which have very different way of prediction. Results of each of them might vary. Determining correct algorithm parameters is an important step here as well.

The last of main objectives is to determine the best way in predicting the results. Predicting draw is very important step, and thus, precision is needed during this step.

Along with these objectives, a hypothesis was made, that day and month the match was played might be important information for its result.

3. METHODS

To achieve the mentioned objectives, a proper dataset is needed. Starting by obtaining only basic information about matches, such as number of goals scored, date played and

result rates and matchweek, the first dataset was built. First dataset contained information about final standing in each of watched seasons as well. Having these data it was possible to count multiple form related and aggregated variables, such as last five matches results, lose and win streaks for last three and five matches, form points, league position, conceded and scored goals in given season, goal and point differences between the teams playing match. These variables represented both team form in given match, according to current season. This dataset is referred as DS1.

Later on, long term statistics were added creating the second dataset, often referred as DS2. More aggregated variables were created in DS2, count from the beginning of whole competition, thus even before oldest data in DS1. This was achieved thanks to all time data available at Premier league official website. Variables such as number of seasons played by both teams up to the given match, total wins, draws and loses up to that match, win to lose ratios and win+draw to lose ratios, conceded and scored goals up to that match, scored and conceded goals per match up to the given match, total played matches, differences between scored and conceded goals, and average standing in all previously attended seasons for both teams were added.

To determine the most fitting number of seasons to train the models a loop with decreasing number of seasons was used. The results were obtained starting off with full number of seasons, 21, and removing the oldest season in dataset (first 380 matches) in each iteration.

Determining the best fitting algorithm was also based by previous loop. Namely, extreme gradient boosting (XGB), logistic regression (LogR), linear regression (LinR), support vector machines (SVM) and random forest (RF). Firstly, the best parameters were determined for each of the tested algorithms using the grid search. After that, all of the algorithms were tested with all of the datasets generated in the loop mentioned above. Best fitting number of seasons was estimated for each algorithm in this loop.

The last test was based on comparing the results of simple prediction model, predicting one of the three possible results, with custom model aiming to predict draw with better success rate. The second prediction model was based on separate prediction of win and something else, and from that lose and draw. Both of these models were tested in previously described loop.

4. SIMULATION RESULTS

First testing was established with the dataset which contained only data about team form. These variables can help in prediction of more possible result, as visible in figure 1. The team position seems to be the most important variable as mostly the team with better league position wins. Teams with better form, more points and more goals scored win more matches as well.

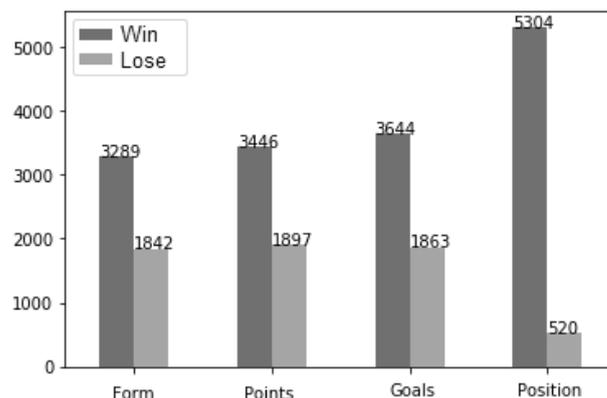


Fig. 1 Wins and losses of a team with better values.

In the second dataset, long term statistics were added to each match. In figure 2, it is clear that these statistics have very similar behaviour as short term statistics. Teams with more goals scored, better average standings, scored to conceded goal differences and better win to lose ratio from long term perspective tend to win more matches. This plot contains aggregated data as we mention later.

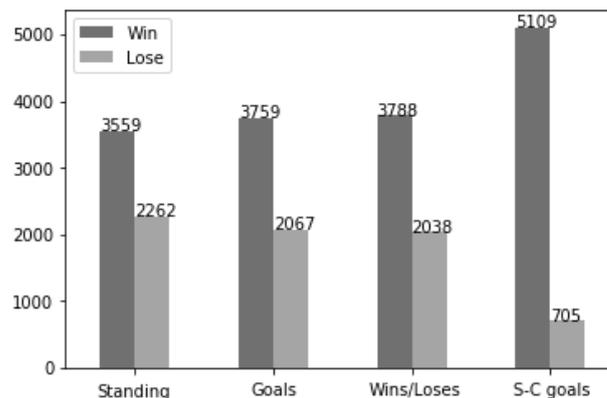


Fig. 2 Wins and losses of a team with better values in long term statistics (better average league standings, more goals, better wins to losses ratio and better scored to conceded goals difference).

At first, only models that determine win or any other result were tested. Testing size was 380 matches, which is one whole season. In table 1, the differences between each dataset are shown. In the second dataset, long term statistic data were used to calculate and add the same value of variables to each match in the dataset. The accuracy rate has dropped in prediction using the second dataset. The cause of this drop is data leakage as there are overall data used in context where they were not known yet. Changing the definition of each variable to be cumulative only up to the match currently playing in dataset 3 came up as a solution. This step had positive impact on the prediction, which can be observable in table 2, where the results of each dataset using 380 matches as testing set and 19-17 seasons as training set are compared. The second dataset led in many cases to prediction rate drop. The third improved dataset corrected this drop and led to better prediction. The only algo-

rithm which didnt improve was SVM, which had very low prediction rate in all cases. During these experiments, date variables were added to each dataset. Separate evaluation was run but no markable positive impact was determined. These variables weren't tested in later experiment for this reason.

Table 1 Comparison of prediction rates of different datasets (1 for dataset with only short term stats, 2 for dataset with long term stats included, 3 for corrected dataset 2) with 21 seasons.

Dataset	LinR	LogR	SVM	XGB	RF
1	0.6711	0.6921	0.6921	0.6868	0.6895
2	0.6658	0.6842	0.6026	0.6816	0.6868
3	0.6868	0.6895	0.6316	0.6842	0.7026
1+date	0.6842	0.6974	0.6868	0.6816	0.6974
2+date	0.6658	0.6868	0.6026	0.6605	0.6816
3+date	0.6868	0.6868	0.6316	0.6868	0.6974

We observed quite increased prediction rates using date information with the first dataset, whereas with the second and third dataset we actually observed decrease in the prediction rates, as seen in table 1. After further tests we decided to drop out date variables from our models.

Table 2 Comparison of prediction rates with different seasons count for each algorithm, with test size of 380 matches and different data sets (DS).

Seasons	LinR	LogR	SVM	XGB	RF
DS1					
19	0.6711	0.6947	0.6868	0.6763	0.7
18	0.6711	0.6947	0.6868	0.6868	0.6974
17	0.6711	0.6974	0.6974	0.6842	0.6947
DS2					
19	0.6605	0.6921	0.6237	0.6868	0.6895
18	0.6658	0.6921	0.6316	0.6684	0.6763
17	0.6763	0.6921	0.6263	0.6737	0.6816
DS3					
19	0.7053	0.7079	0.6289	0.6895	0.6868
18	0.7105	0.7132	0.6316	0.6842	0.6974
17	0.7053	0.7	0.6316	0.6868	0.7026

Within the previous tests, number of seasons best to use to train the models was analyzed. A different season count

with test size 380 matches is compared in table 3, where in table 4 testing set was 70 matches. Modular season count was achieved by dropping out older seasons from dataset. The results are clearly not the same and some season counts produce better results. The modular number of seasons in training set had different impact on each algorithm. The impact was also different when testing set was 380 and 70 matches. This might be caused by fact, that the difference is almost one whole season. In case of test size 70 matches, there is greater error in prediction than with larger testing set. With 380 testing matches, the best results achieved by regression algorithms were with 18 seasons, random forest had best results with 15 seasons, XGB had best rate with 20 seasons and SVM had about the same results with any dataset length.

Table 3 Comparison of prediction rates with different seasons count for each algorithm, with test size of 380 matches.

Seasons	LinR	LogR	SVM	XGB	RF
21	0.6868	0.6895	0.6316	0.6842	0.7026
20	0.6868	0.7	0.6342	0.6974	0.6947
19	0.7053	0.7079	0.6289	0.6895	0.6868
18	0.7105	0.7132	0.6316	0.6842	0.6974
17	0.7053	0.7	0.6316	0.6868	0.7026
16	0.7	0.6947	0.6342	0.6763	0.7026
15	0.6921	0.6974	0.6395	0.6789	0.7105
14	0.6895	0.6974	0.6395	0.6763	0.7026

Table 4 Comparison of prediction rates with different seasons count for each algorithm, with test size of 70 matches.

Seasons	LinR	LogR	SVM	XGB	RF
21	0.6571	0.6429	0.5857	0.6429	0.6571
20	0.6714	0.6714	0.6	0.6571	0.6714
19	0.6857	0.6714	0.6	0.6429	0.6429
18	0.6857	0.6857	0.6	0.6286	0.6429
17	0.6571	0.7	0.5857	0.6429	0.6714
16	0.6857	0.7	0.6143	0.6143	0.6143
15	0.6857	0.7	0.6143	0.6571	0.6714
14	0.6714	0.6714	0.6143	0.6286	0.6714

Along with determining the best fitting dataset, searching for the best parameters using the grid search was applied. Test size used during grid search was 380 matches. Both regression algorithms achieved the best prediction rate with default parameters, 71% in both cases. Better

progress was achieved with SVM algorithm. First, the kernel was changed from rbf, but both polynomial and linear kernels had worse results as rbf. After tuning other parameters of SVM, the prediction rate increased from 57,14% to 63,16%. For XGB, the default parameters were the best, where the rate reached 69%. However, the best progress was achieved with random forest algorithm parameter tuning, where the rate increased from 62,86% to 70%.

In the last step of tests, algorithm to predict win, draw and lose in separate steps was designed. Hypothesis was set to increase success rate of draw prediction using such approach. It is quite easy to determine the winning team according to results of the previous test. Models can as well predict the probability of any result. The probability that the home team will win was predicted by the model we tested up to now. The model was labeled WX. After that, the second model (named XL) was used. XL predicts only if the team will lose. Adding up probabilities together, all three possible results probabilities can be determined. This algorithm was compared with the results from the third model, which predicts one of the three values - win, draw or lose in one step.

In tests described up to now, only win probability was predicted. However, the simpler model was evaluated firstly. This model predicts exact result. Linear regression was due to the nature of the algorithm excluded from this test. Results can be seen in table 5. Algorithm SVM achieved the lowest score from tested algorithms, 54%. The rest three algorithms had average scores 59,18% for logistic regression, 59,67% for RF and 60,2% for XGB. The best score achieved was 61,05% with 21 seasons using XGB. Differences between tests with variable dataset length were not crucial in this case.

Table 5 Comparison of prediction rates with different seasons count for each algorithm predicting exact result in one step, using 380 matches for testing.

Seasons	LogR	SVM	XGB	RF
21	0.5868	0.5263	0.6105	0.5974
20	0.5947	0.5316	0.5974	0.6026
19	0.5974	0.5368	0.6026	0.6053
18	0.6	0.5368	0.6026	0.6
17	0.5947	0.5395	0.5974	0.6026
16	0.5974	0.5684	0.6	0.5947
15	0.5895	0.5474	0.5974	0.5895
14	0.5737	0.5526	0.6079	0.5816
Average	0.5918	0.5424	0.6020	0.5967

The designed algorithm was then evaluated to determine, if we can get better results. According to table 6, the best results are in range 53-55%, where averages in range 43-51%. The cause of these results seems to be unbalanced

dataset, where home team win appeared in 46% of matches, but loss only in 25,5% cases, which means the XL dataset had 74,5 to 25,5 split.

Table 6 Comparison of prediction rates with different seasons count for each algorithm predicting results by models WX and XL, using 380 matches for testing.

Seasons	LinR	LogR	SVM	XGB	RF
21	0.4811	0.4872	0.4573	0.5031	0.4648
20	0.4555	0.4644	0.4156	0.4803	0.4742
19	0.5235	0.5334	0.4730	0.5316	0.5264
18	0.5320	0.5496	0.4342	0.5136	0.5091
17	0.5336	0.5107	0.4480	0.5373	0.5060
16	0.4738	0.4661	0.4242	0.4729	0.5202
15	0.4826	0.5397	0.4355	0.5480	0.5040
14	0.4539	0.5199	0.4088	0.5322	0.5111
Average	0.4920	0.5089	0.4371	0.5149	0.5020

In the direct prediction, model predicted draw 108 times out of 380 matches, where 71 were correct. This means that the direct prediction method can predict draw without problems. These results are quite surprising, comparing to results of mentioned papers. This also means, that we managed to achieve better results than results of data we used while training.

5. RESULTS

Even by using the basic data obtainable freely on the internet it is possible to generate comparable results in sport result prediction. Using only short term statistics and in several papers mentioned in introduction doesn't seem to be the best idea. Results were clearly better with long term statistics included. It seems that the algorithm based on hypothesis that separate prediction might increase prediction rate wasn't the best possible way.

Even after parameter tuning of SVM algorithm, the results reported by this algorithm were markedly lower than other algorithms. Linear regression might not be useful for this problem, as it can only determine binary problems. From the three algorithms left, each had better results compared to the papers mentioned. Best results were achieved by the algorithm XGB, over 60% in average. Results of logistic regression and random forest are close to this value, with 59,18% and 59,67%. With the designed algorithm, the best prediction rates were 51,49% for XGB on average, which is almost 10% lower than with direct prediction. Other algorithms had results close to 50% or even less.

6. DISCUSSION/CONCLUSIONS

A dataset was created with limited sources and limited number of rows. The results achieved by training models

with this dataset are comparable with previously published papers.

One of main goals of this work, to increase success rate for draw prediction was successfully fulfilled. The draw prediction rate on testing set was 65%. Out of 380 matches, 108 were predicted as draw, where 70 were correct. Thus the draw prediction rate is higher than any of the investigated works.

Three of the algorithms used within this paper had results comparable with other works. Namely, XGB with average accuracy 60,2%, logistic regression with 59,18% and random forest with 59,67%. Long term statistics led to better results than only short term results.

The hypothesis that split prediction might increase accuracy has turned out incorrect. So has the hypothesis that date might help.

Support vector machines ended up with low prediction

rate and thus seem to be unsuitable algorithm for this problem.

Including more information about the matches such as player statistics, wheater conditions and team tactics could be helpful. An interesting idea might be social network analysis, watching how people act about a given match. In case of betting system, a number of bets on each result can be included in prediction, making prediction more dynamic. By now, only static data were used.

ACKNOWLEDGEMENT

This work was supported by the Slovak Research and Development Agency, project number APVV-18-0214 and by the Scientific Grant Agency of the Ministry of Education, science, research and sport of the Slovak Republic under the contract: 1/0268/19.

Received September 25, 2019, accepted November 30, 2019

REFERENCES

- [1] PURUCKER, M. C.: Neural network quarterbacking, *IEEE Potentials* 15.3, 1996, 9–15. DOI: 10.1109/45.535226
- [2] KAHN, J.: Neural Network Prediction of NFL Football Games, *World Wide Web electronic publication* 2003, 9–15. <http://homepages.cae.wisc.edu/~ece539/project/f03/kahn.pdf>
- [3] MCCABE, A. – TREVATHAN, J.: Artificial Intelligence in Sports Prediction. *Fifth International Conference on Information Technology: New Generations*, 2008, 1194–1197. DOI 10.1109/itng.2008.203
- [4] TAX, N. – JOUSTRA, Y.: Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. *Transactions on Knowledge and Data Engineering* 10.10, 2015, 1-13. DOI 10.13140/RG.2.1.1383.4729
- [5] KUMAR, G.: *Machine Learning for Soccer Analytics*, 2013, DOI: 10.13140/RG.2.1.4628.3761

BIOGRAPHIES

Martin Chovanec received his Engineering degree in Informatics in 2005 from the Faculty of Electrical Engineering and Informatics, Technical University of Košice. In 2008 he received his Ph.D. degree at the Department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics of the Technical University of Košice and his scientific research was focused on network security and encryption algorithms. Currently, he is Director of the Institute of Computer Technology of the Technical University of Košice.

Marek Ružička was born on 1. 1. 1995. In 2017 he graduated Bc. and in 2019 Ing. with distinction at the department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics at Technical University in Košice. He now studies as a PhD. student at the same university.