The application of machine learning principles in sports betting systems

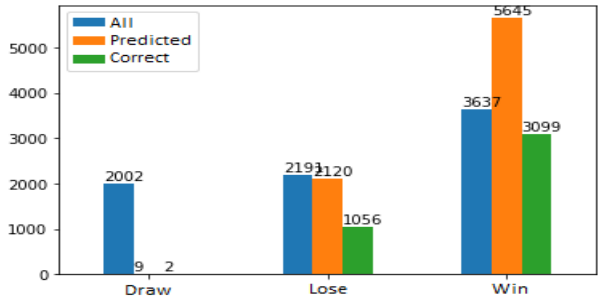
Marek Ružička, Juraj Gazda
Department of Computers and Informatics, Technical University of Košice, Slovakia

MOTIVATION

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. The most recent works for soccer prediction report success rate in range 53-56% [1][2][3]. Draw prediction seems to be a crucial problem in the field of result prediction.

ANALYZED OBJECTIVES

- Usage of short term statistics compared to long term statistics for 21 seasons of Premier League matches.
- Comparison of five different machine learning algorithms in order to get the best possible prediction rate. The compared algorithms were linear regression, logistic regression, support vector machines (SVM), gradient boosting (XGB) and random forest (RF) implemented in python.
- Using different number of seasons to determine, if the old data could possibly decrease success rate.
- Comparison of direct prediction of three possible result to splitting the prediction into two binary problems.



Draws, loses and wins within 21 seasons – all, predicted and correct in retrieved dataset

METHODS

- Multiple datasets were used to determine whether is it better to use short term information only – mostly current team form – or long term information as well, such as average number of goals, win to lose ratio and so on. All datasets were tested with and without date variables.
- Dataset was periodically cut by oldest season and models were re-trained with smaller datasets to determine, how many seasons in the dataset produce the best results.
- Dataset where all three possible results were used for prediction was compared to the dataset where we first tried to predict if the home team will win or not and after that if it will lose or not.

RESULTS AND CONCLUSION

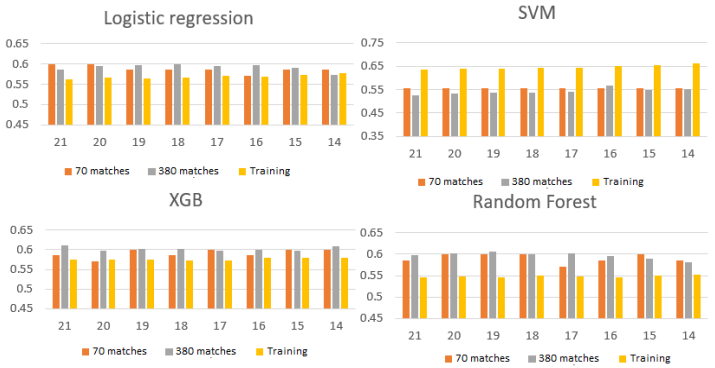
From the simulation results we can conclude several conclusions for the objectives.

Long term statistics seem to have positive impact on prediction rate in most cases, however, its impact is not that great as we would expect. The biggest impact was cased by variables related to the points gained in recent matches. Another very important variables were rates given by external betting company. This rate was used as replacement of data about bets we would have if we run real betting system where bets are placed by customers.

The variance in prediction with different dataset length was more noticeable with testing dataset where we tried to predict only 70 matches. In case we tried to predict 380 matches, there were minimal changes using different dataset lengths. The success rate varied by 1% in most cases. It seems that there is higher chance for error in case we use smaller dataset for validation.

With the split prediction of win and lose we haven't achieved any good result. Success rate was in the range of 51-55%, where in many cases it didn't pass 50%.

We have achieved very good results with direct prediction. Algorithms logistic regression, XGB and RF all reached success rate of 59-60% with 380 matches in testing dataset. Algorithm linear regression didn't pass 40% in any case and algorithm SVM passed 50% in just one case. The best result was achieved with XGB, exactly 61.05%.



Results of four different algorithms using variable length of training dataset with test set of 70 and 380 matches

These results passed any expectations and are good enough to be used in real betting system.

References

[1] Tax, N., Joustra, Y (2015) Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach DOI: 10.13140/RG.2.1.1383.4729
[2] Kumar, Gunjan (2013), Machine Learning for Soccer Analytics DOI: 10.13140/RG.2.1.4628.3761
[3] http://kickoff.ai (Jan 2019)