Minorproject 2017

Introduction

Every year, more money is spent on professional soccer players and more bets are being put onto those players and their teams. Therefore it could be profitable to find a way to predict the outcome of a soccer match. When this is done properly, the world of soccer will change and teams will probably base their purchases on this prediction model. Interesting input values for the model could be a players physical performance. When the average physical stats of a team are known, maybe this can predict the winner accurately. Another interesting kind of data would be the odds of betting companies, or the skills of the players of a team according to the game Fifa, the biggest soccer game franchise existing. When the data of almost 25000 matches is used to predict the outcome of these matches, maybe soccer can be made predictable.

In this paper, a way to predict the outcome of soccer matches is tried to be found using a “kaggle” dataset consisting of 25000 soccer matches played from 2008 until 2016. Stats from matches played in 11 different leagues are known, including the 22 players who played in a specific match, the stats of all of these players, updated twice every season and the identity of a team expressed in 8 values. Also the betting odds of 10 different betting companies are included which represent the profit made when betting on the outcome of a match.

Using the predictors above, a prediction is being constructed using different algorithms like an Artificial Neural Network and a Bayesian Network. Then based on this prediction, the total profit made after betting on a number of matches, according to the predictions, is being calculated.

“Goal” of the project

To have a clear view of what has to be produced, a goal is being determined as following:

**“Making profit when betting on soccer matches based on a prediction of the outcome which is calculated using stats from the game Fifa and the odds of betting companies”**

It is very hard to state how much profit is realisable, which is why the goals just says “profit”, which means anything above zero. The reason profit is being chosen as a goal, over for example accuracy, is because of the complexity of the prediction. Normally, a problem that has three possible outcomes (win, draw loss), would be predictable with at least an accuracy of 33.3% (1/3). In this case, this percentage is different. The “home-advantage” causes an increase in the minimal accuracy. When all the matches have “home wins” as prediction, an accuracy of 47% is being achieved. Because of this, the profit will be the output variable that will be measured.

“Sub-goals”

Because the main goal is very generic, in this chapter, some sub-goals are stated.

* Finding the best predictable league
* Finding the most influential player in the team
* Finding the most influential team stat
* Finding the betting company that generates the most profit

When these sub-goals are reached, the amount of profit that can be achieved will be maximized.

Data processing

Before a prediction model can be used to predict the outcome of soccer matches, the available data needs to be sharpened to get ready for usage. This means removing outliers, joining tables and calculating new variables.

After removing entries that miss either the outcome of a match, any of the players, the date or any of the team identifiers. The tables “match”, “player attributes“, “league” and “team attributes” are joined in a way that the following list of variables in present in the same table. This has been executed in an sqlite editor.

(Match\_id, date, H\_pid1, H\_rating1, H\_po1, H\_team\_id, H\_buildUpPlaySpeed, H\_buildUpPlayDribbling, H\_buildUpPlayPassing, H\_chanceCreationPassing, H\_chanceCreationCrossing, H\_chanceCreationShooting, H\_defencePressure, H\_defenceAggression, B365H, B365D, B365A)

In this table, the Match\_id serves as primary key, for this is a unique value. The date is necessary to calculate the correct ratings of the players. Notice that for each player in the home and away team, the player id, the rating and the potential (pid, rating, po) are in the table (66 values), but are not in the list above, for readability.

<https://www.ncbi.nlm.nih.gov/pubmed/16195009>

<http://www.tandfonline.com/doi/abs/10.1080/0264041031000071182>

<https://www.researchgate.net/profile/Samuele_Marcora/publication/6769681_Validity_of_Simple_Field_Tests_as_Indicators_of_Match-Related_Physical_Performance_in_Top-Level_Professional_Soccer_Players/links/02e7e52fe96b6da444000000.pdf>

<https://www.kaggle.com/hugomathien/soccer>

Different people have tried to declare what makes a player “good”. Mohr, (2003) focusses on the fatigue of players during a match, while Rampinini, (2007) has investigated the physical components of players to predict their performance. Svensson and Drust, (2005) claim that any measurement of players attributes can’t be used to predict player performance in a match. The complex nature of performance in competition makes this an impossible value to calculate.

However, what these publications have in common, is that they use data only about the players performance.