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.

This table will function as the base for all algorithms that will be carried out on the data in python. An interesting feature about the data is that it is very well up to date. As there are 8 years of data, players’ ratings change over time. For example Cristiano Ronaldo wasn’t as good 8 years ago as he is now. Both the players’ individual ratings as the team ratings are retrieved on a date nearest to the date of the match. So when a certain match was played on 16-06-2017, the ratings are those retrieved on the closest date before this date. This feature makes the data more trustworthy and probably better predictors.

Data about the data

To get a clear view of the processed data, in this chapter a summary about the data is being presented. Some interesting features are also revealed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| League | #Matches: | Home win | Draw | Away win |
| Belgium Jupiler League | 1215 | 46,7% | 24,7% | 28,6% |
| England Premier League | 2962 | 45,5% | 25,9% | 28,6% |
| France Ligue 1 | 2864 | 45,0% | 28,1% | 26,9% |
| Germany Bundesliga 1 | 2376 | 45,4% | 24,4% | 30,3% |
| Italy Serie A | 2747 | 46,6% | 26,3% | 27,1% |
| Netherlands Eredivisie | 2035 | 47,3% | 23,5% | 29,1% |
| Poland Ekstraklasa | 444 | 46,6% | 27,7% | 25,7% |
| Portugal Liga ZON Sagres | 1198 | 44,1% | 25,9% | 30,1% |
| Scotland Premier League | 1541 | 42,5% | 23,7% | 33,8% |
| Spain LIGA BBVA | 2707 | 48,5% | 23,4% | 28,1% |
| Switzerland Super League | 1157 | 45,3% | 24,3% | 30,4% |
| **Total** | **21246** | **45,9%** | **25,2%** | **28,9%** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Season | #Matches: | Home win | Draw | Away win |
| 2008/2009 | 1763 | 46,8% | 25,2% | 28,0% |
| 2009/2010 | 2462 | 48,3% | 24,8% | 26,9% |
| 2010/2011 | 2674 | 46,1% | 26,1% | 27,9% |
| 2011/2012 | 2778 | 46,7% | 25,2% | 28,1% |
| 2012/2013 | 2880 | 45,0% | 25,7% | 29,3% |
| 2013/2014 | 2777 | 46,4% | 24,2% | 29,4% |
| 2014/2015 | 2981 | 44,8% | 25,3% | 29,9% |
| 2015/2016 | 2931 | 44,0% | 25,3% | 30,6% |
| **Total** | **21246** | **45,9%** | **25,2%** | **28,9%** |

Prediction strategies

The first thing that comes to mind when trying to predict the outcome of a soccer match, is the skills of the players in a team. The team that has the best players will win the match, or at least has the biggest chance of winning a match. 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. This would mean the players’ skills can’t be used as predictors at all. However, what these publications have in common, is that they use data only about the players performance. Maybe other attributes of a team are more influential when it comes to winning a match or losing it. The game Fifa, created by EA has ratings for all players in the game, and some characteristics about the teams as wholes. These values might be could predictors.

Another type of predictors are the odds that betting companies produce before a match is played. These odds are based on the bets of betters, and earlier played matches. These values can be used as input to an algorithm that predicts the outcome of a match. The goal here would be finding matches that the betting companies have “predicted” wrongly, but an algorithm that uses both the Fifa data as the betting odds as input can actually predict. When this is the case for a match, profit can be made. Let’s look at an example. When team A plays against team B, the odds of a betting company are 6 a home team victory, 4 for a draw and 1.53 for an away victory. This indicates a prediction that the away team will win by the betting company. When the algorithm predicts a win for the home team, which turns out to be correct, this could result in high profit margins.

Betting strategies

The goal of this project is to make profit in betting on the outcome of a football match. Before any algorithms are used in this, some standard betting strategies are being tried out.

**Home team always wins.** It turns out that in many sports, the home team seems to have an advantage, if this advantage is great enough, it could be a good strategy to always bet on the home team.

**Always bet on the lowest betting odd outcome.** Maybe the prediction made by the betting company is that good, that following their prediction is the best strategy.

**Random bet.** It seems to be very hard to predict the outcome of a football match, which might mean that it is so random that random betting is the best strategy.

**Betting based on the prediction of an algorithm.** Probably the best strategy to use, as it takes all the other strategies into consideration and uses the data that has the most influence to predict.

Restrictions

When the most profitable strategy has been found, it is important to look at restrictions to that strategy. As matches have more attributes than the ones used in the algorithm prediction, it might turn out that the matches in a certain league are easier to predict than those in another league. Also it might be the case that matches played in the end of the season are easier targets, or that matches get more predictable every year. A third option would be to combine the strategies, only betting on a match if the prediction of the algorithm says that the home team wins. In short the dimensions that can be explored are time, league and the outcome prediction.

Predicting the outcome

Now that it is clear what has to be explored exactly, these strategies can be tried on the created data table. The first step is to turn the number of goals of both the home and away team into one variable, which has a value of 1 (home team wins), 2 (draw) or 3 (away team wins). Now predictions can be made. For every match in the database, an output of one of these three numbers has to be generated, so that it can be compared to the real outcome and some prediction measurements can be found. When this is done for all of the strategies, the best strategy can be chosen and the restrictions can be tried out to tweak the strategy. When it turns out the best strategy seems to be the use of some prediction algorithm is the best strategy, the parameters of this algorithm can be tweaked as well.`

Measuring the prediction

There are several ways to measure if a prediction was successful or not. In this chapter the measures that will be used are established, along with an explanation. Next to the measurements that usually take place within predictions, this prediction problem has a twist. In the end the goal is to make profit out of betting on soccer matches, so that means profit on its own is a plausible metric. Also, even when the usual metrics seem to be high, this doesn’t mean the profit will be high. It is important that the profitable matches are being predicted well, and not so much that as many matches as possible are predicted well.

More in depth, this profit comes from a number of matches, so maybe the profit per match would be a better choice, which leads to profit margin per match as the input bet can of course alter the profit. Imagine a profit margin of €10,- that comes from 2 matches and the same profit coming from 4 matches. With a budget of €4,-, in the first case a profit of €20,- can be achieved, while the second case only establishes a profit of €10,-. However, when all the profit comes from a few matches, this comes along with a high risk. In future predictions, this 1 match could turn out to be wrongly predicted. This wouldn’t be as much of a problem if the budget was divided over more matches. The balance between profit per match and risk needs to be found.

The prediction algorithms

As there are lots of algorithms that can be used to produce a prediction, a selection is made here based on what is normally used in predictions in sports. With every algorithm comes a short explanation and the parameters which are important to keep in mind.

PREDICTION ALGORITHMS

<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>