Football Datasets

Jonas Stillhard

Swiss Federal Research Institute WSL
jonas.stillhard@wsl.ch

Motivation

- Worldcup 2018 in Russia
- Focus on Football
- Potential for a widely received publication / blog post

Available Datasets

- **Openfootball aka football.db:** Available on github. Can be downloaded to a sqLite database via command prompt using ruby. Very convenient. Provides scores for all fixtures since 1930.
- **footballdata.uk:** Provides final scores for 10 european countries and 22 leagues including betting odds. Updated weekly, data is available as .csv, data goes back to 1994. Very interesting dataset, might be useful for other analyses.
- **engsoccerdata:** Historical football datasets with a focus on english premier league, historical data back to 1871. Available on github.
- Check out **Joe Kampschmidts** blog for more datasets and API's.

Restrictions

- Predictions of final scores and winners of tournaments integrating as much predictors/variables as possible are very common - e.g., Bloomberg tried to predict the winner in 2014.
- Data availability is not restricted if using more general data (e.g., final scores). 'Fancier' data (e.g., position of players when scoring goals) are not available.

Ideas?

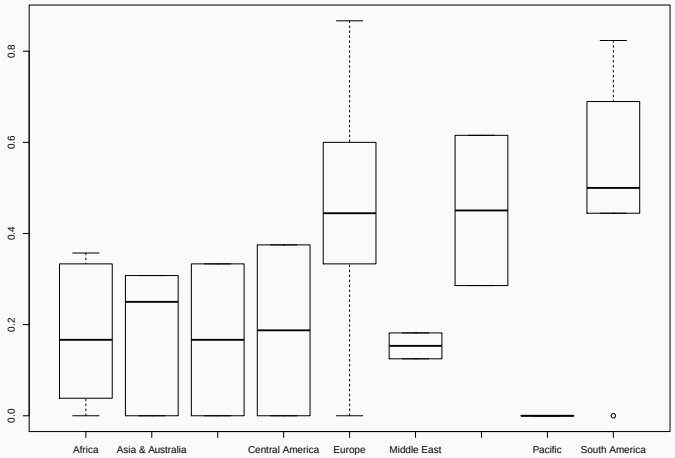
Preliminary ideas

- Use available datasets.
- The goal is **not** to outperform existing models
- Provide rules of thumb for teams that win and make these rules as easy as possible (e.g., countries starting with A, B, C, D, E, F will always win if playing against another country).
- Benchmarking: Use real world data if available (we might get a world cup 2014 office pool dataset from Switzerland (N¿1000)) or betting odds.

'No clue, here's what you do'

- Provide people not interested in football with good rules of thumb so they can perform all right in an office pool.
- What strategies are most promising for different types of betting pools?
- Depending on the amount of time one is willing to invest for an (office) pool, what is the best strategy (e.g., if you are willing to have a look at the betting odds every day, follow these, else always bet on a 1:0).
- The main goal is not to win money but to maximize social standing with a minimum effort.

Preliminary analyses



Data

The data is available on github in a private project, do not hesitate to request access. So far, the dataset contains all world-cup games until 2010. Due to a bug, the data of the 2014 world cup can not be downloaded at the moment.

`https://github.com/jstillh/ftblData.git`

Questions?