Plan Overview

A Data Management Plan created using DMPonline

Title: Analysis of correspondence between real-world football players' statistics and FIFA players' values and rating

Creator: Ivan Lichner

Principal Investigator: Ivan Lichner

Data Manager: Ivan Lichner

Project Administrator: Ivan Lichner

Affiliation: Other

Template: DCC Template

ORCID iD: 0000-0001-8473-6573

Project abstract:

In this project I will collect the statistical data from the 5 biggest European football league - namely Premier League, La Liga, Serie A, Bundesliga and Ligue 1 as well as the data from the game FIFA. The statistics will be for the season 2016/2017 and the equivalent FIFA 18. I will then calculate the correspondence of the real performance of the player on the field with the FIFA value and overall rating of the player. The results could be used to gauge to which extent the developers of the game take into account the players' real performance or the fame of the players is more important when determining the value and rating for the game. I will use my own scoring weights for various statistics. At the end I will list the worst judged players in the FIFA game.

ID: 75069

Start date: 01-04-2021

End date: 18-04-2021

Last modified: 24-04-2021

Analysis of correspondence between real-world football players' statistics and FIFA players' values and rating

Data Collection

What data will you collect or create?

I will collect three types of statistical data for each league - namely standard statistics, goalkeeping statistics and miscellaneous statistics, and merge them together and then select the usable parameters. These statistics are already collected and I downloaded them from the webpage FBref in the comma-separated value format. I use this format, because of its standardization, widespread usage and compatibility. Standard and miscellaneous data-sets for each league contain data for all players in the league and are bigger than the goalkeeping data, but all together for each league I need 111 - 134 kB of space.

Additionally, I will use the FIFA data-set from KaggleFIFA18 which is also in CSV format. This data-set has size 18,6 MB.

I will create a new data-set for each league with the calculated statistical data based on custom weights and also the selected parameters. My created data-sets have the size 24 - 29 kB.

How will the data be collected or created?

The FIFA18 data-set from Kaggle is downloaded, renamed to FIFA18_complete.csv and loaded into notebook.

The real-world statistics of the players are exported on the page FBref, then copy-pasted into the corresponding file - standard stats are stored in files ending with suffix _STD, goalkeeping stats with suffix _GK and miscellaneous stats with suffix _MISC. Each league has its own prefix in the file name - Premier League has PL, La Lila has LL, Serie A has SA, Bundesliga has BL and Ligue 1 has L1, e.g. standard statistics file for Premier League has a name PL_1617_STD.csv. After the loading, first comes the elimination of duplicates in statistics which is only in case of transfer of player during the season and then the data-sets for each league are merged on player's name, and we have obtained one big data-set which collects all three types of statistics. The elimination of duplicates consists of summing the parameters of duplicated players to preserve the performance during the whole season.

Now follows the selection of the appropriate parameters for my experiment from this big data-set and renaming the columns to human-readable and understandable form. I have chosen the parameters that characterize the performance of each position of a player the best and created a weighting strategy for each parameter. I considered the positions Attacker, Midfielder, Defender and Goalkeeper and the weights are assigned in such a way that each position benefits from the main responsibilities that it has.

After parameter selection and deciding on the weighting strategy I calculated the real-world statistics of the players and joined the results with the FIFA 18 data-set to perform the correlation analysis as well as listings and visualization of the distribution of players' values for each league.

The data-sets files are stored in one folder, with the mentioned naming convention for each league.

In my experiment the versioning is not considered since we are using the data for the season 2016/2017 and FIFA 18, which were collected and refined many times before I used them and are not updated anymore.

Documentation and Metadata

What documentation and metadata will accompany the data?

I will use the metadata format defined by the Dublin Core Metadata Initiative standard, and it will consist of mainly author information, topic description, data-sets used and rights and dates. The metadata will be stored in the GitHub repo in the directory **metadata**.

Ethics and Legal Compliance

How will you manage any ethical issues?

The data-sets are free to download and use under CC0 forFIFA18 and under general law that the facts cannot be copyrighted for statistics from FBref.

With regards to the sensitivity of the data used, there are none. I am using the real statistics from the open leagues which can be

collected by anyone who is able to read or hear so there are no secrets revealed.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The owners of the data are Sports Reference LLC for the statistics and the FIFA 18 data-set is shared under CC0 and so there are "no rights reserved" for the owner anymore.
If any IPR issues occur in the future with regards to the collected statistics, I can always handle them on the fly.

Storage and Backup

How will the data be stored and backed up during the research?

Since this is a small project and I will also work alone on it, I will store the data-sets files at local computer, in a private Dropbox account as well as GitHub repo. The source code of the Jupyter Notebook that does the loading, calculations, listings and visualization is stored on GitHub in an open repository as well as on a local machine.
The whole space required for the experiment is less than 20 MB, so no special storage capacity is required.
In case that any of the data-set files or Notebook file went missing I will reload them from the secondary/tertiary storage (local machine).

How will you manage access and security?

Since it is requested that this is an open experiment and as mention in the previous question I am using an openGitHub repository, the access to the Jupyter file is free to anyone to use and share it.
The security of the stored data is provided by the GitHub and Dropbox security policies and since both of them are companies with broad user base, they have already implemented the sufficient security strategies. Both of them also provide SSH link to the resource for the possible new collaborators. On my private laptop I am using an antivirus program, that should ensure the safety and security of the files.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Since we can collect the data whenever we want there is no need to retain any of them, but for the safety of reproducibility of my experiment I will store the downloaded data-sets on GitHub server, which will also ensure its availability indefinitely.
The source code of the Jupyter Notebook will be preserved byGitHub and its policies say that it is available indefinitely.

What is the long-term preservation plan for the dataset?

The preserving of Jupyter file in GitHub repository as well as the data-sets is free of charge if it's within the GitHub limits for free repositories.
Since both Jupyter file and the data-sets use the standardized formats, there is no cost of preparing them for sharing or preservation.

Data Sharing

How will you share the data?

I will share the DMP on Zenodo platform with a publication status and also the whole experiment is stored in aGitHub open repository, so the data are there. The data will be available from the 15.04.2021, since the experiment started at 01.04.2021.
Regarding the reuse I will require to copy the link to the original repository of the experiment.

Are any restrictions on data sharing required?

Since the FIFA dataset has been published with a CC0 license there are no restrictions regarding the usability of the data and the real-world statistics are the open data, so there is no restriction there as well, I only need to cite the page where I have got the data from.

Responsibilities and Resources

Who will be responsible for data management?

Since I am working alone on this project with no future collaborators in mind, I am responsible for all the duties reagrding this experiment be it writing and revising of the DMP, data management activities or metadata production.

What resources will you require to deliver your plan?

I am using the free open-source application Jupyter Notebook which is available for the download on any operating system so there are no special requirements on SW or HW for my experiment. The data-sets are not so big that any additional processing power than the ordinary personal computer has available is necessary.