**Football Analytics: A Case Study on the Availability and Limitations of Data for Football Analytics Research**

Aim of the paper:

The main objective is to develop a player rating system - a system capable of calculating player rating based on the values of the events recorded in the dataset.

Dataset Used:

On 17th August 2012 OPTA publicly released a rich dataset of player performance statistics of English Premier League season 2011-2012. It contains frequency counts for hundreds of different events, ranging from aerial duels to yellow cards, broken down by player and by game. This dataset contains close to 10369 rows of data and approximately 210 attributes.

Statistical measures/ validation approach used:

The Document Object Model (DOM) was observed and it was observed that it was, in fact, not a

static website but displayed data through AJAX requests. The heavy use of Javascript in displaying match data made it very hard to scrape unless we could decipher a link to the JSON data. There was a need to create an environment that would mimic a user accessing the website through a browser. This is when PhantomJS was come across, a truly headless Webkit browser that runs JavaScript code, and has no dependencies.

Once the complete data was gathered it had to be converted from the JSON format and incorporate it into the dataset which was a csv file. This is the next test that was encountered.

Accuracy/error:

The consistency and of the data needed to be checked; it was noticed that some players represented by their nicknames in either of the two sources from where the data was collected example (Javier Hernandez was named as Chicharito) and some teams with larger names were represented by their shortened names (Wolverhampton Wanderers as Wolves). These discrepancies called for the data to be cleaned.

Conclusion:

Through the course of the development of this study we observe that although there is abundance of data on football to perform operations of varying degree, the nature and the detail that data covers vary from source to source. Information and data relating to the result and outcome of match including betting data is readily available, however when it comes to working on player performance evaluation and player in-game events statistics there is a lack of publicly available data and the presence of high-end data collection and analytics companies working in the field of football analytics developing their own datasets for use by other interested parties lead to a complicated situation. The method of data collection used in this case study outlines the combination of a common trick that is used for gathering data from websites that do not provide access through API’s and applies a customized solution of gathering the desired data. The collected data from whoscored.com acts as the second data source therefore the two datasets needed to be combined, it is here that some common ground between datasets was needed which led to the need for data integration and cleaning operation to be performed.

Why this is important:

Data is equivalent to gold today, and the lack of publicly available data on individual players is truly horrible. Publicly available data benefits everyone and can help in a tremendous manner in teams trying to recruit new players into their teams.

This problem is aggravated by the fact that several companies who are working on collecting data for this purpose create their own datasets with their own formats and their own parameters, which makes working across the data from companies impossible, assuming we can even get data from multiple companies.

There is also a lack of incentive to create a public database of players and their attributes, as companies charge a hefty sum for their services.

This paper is important to create awareness among fans and team managers to lobby the FIFA association to publicize data, which will help tremendously improve the current situation. And, with the amount of data that’s present all over, the findings, I’m sure, will be very very interesting.

My critiques:

My biggest critique of this paper is the methodology they have used to collect the data. The method used can be described as illegal, as it exploits a method not accepted by the companies the data is taken from.

I strongly condone illegal practices to obtain data, and also hope we can find a way to obtain the invaluable data from companies.

Done by-

Siddarth S (PES2UG20CS336)