Automated football pre-match analysis using python-docx.

Alessandro Rossi (4641902)

Objectives

- Analyse the issues of building a pre-match report, its methods and the best solutions.
- Analyse the issues of communication between the various figures in a football team (Analyst, Coach, Players, Management).
- Develop a python script to automate this process.
- Utilise at least one uncommon and/or innovative method.

Introduction

To work on this project I will be using a simplified demo data provided by Wyscout Hudl, for simplicity I will also refer to Team A as the team doing the report and Team B as the opposing team.

All the code and examples will be available in this repo.

While sports have existed for hundreds of years, professional football was legalised in England in 1885, rigorous data analytics are a modern addition.[1]

It started in 1996 in the UK when a new company called Opta partnered with the Premier League to build infrastructure and accessible data to the football clubs, by using cameras and automated tools. This model was quickly implemented in every other major league, and it is the base of modern analytics.

This however created a new problem, where some teams, especially in England, preferred a data-driven approach while some teams preferred an “old style” approach where the Coaches and players apply their game knowledge to decide strategy. This is not done out of spite or anything like that, but it is simply due to a difficulty of communication in different fields, coaches for the most part are ex-players while analysts are experts in the data science field.

A possible solution to this problem is clear, yet powerful data that can be used by both these figures.

Data analytics

What variables to take into consideration

Choosing what variables to take into consideration is a difficult choice. It is difficult because most of the time correlation does not mean causation, a lot of these stats lack the context of the games themselves. Another issue is that the sample size is not big enough which makes any outlier have a big impact in the final result, it is mostly pointless to take into consideration older stats as Team B might have had a completely different line-up just a single competition ago, in this case I take a sample size of 14 games which is going to be unreliable.

Possible ways to reduce this uncertainty are to rely on the knowledge and expertise of the coaching staff and not to take into consideration the outliers and give these stats proper context. Another way is to only take into consideration the most impactful and consistent stats, for example number of passes tends to be a good stat as there can be hundred of passes in a single match, which adds up to thousands over many games and is unlikely to be affected by a “bad” game. For the purpose of this paper, I will be relying on these two papers:

[2][3]

As it is probably known, shots related variables tend to be the most reliable, total shots on target from the penalty box is a very good stat. Number of Accelerations also tend to correlate to more successful teams.

A stat that surprisingly correlates with worse results is Number of Crosses.

Introduction to python-docx

A tool I have discovered while working on this project is python-docx, an open-source python library, it is a powerful tool that gives an analyst the possibility to make automated docx reports that are highly customizable, modular and easy to read for all, coaches, analysts, players etc.

The report will start with raw stats, it must be able to fetch stats in any form. It will include a stat-sheet, a section of stat manipulation and finally some conclusions.

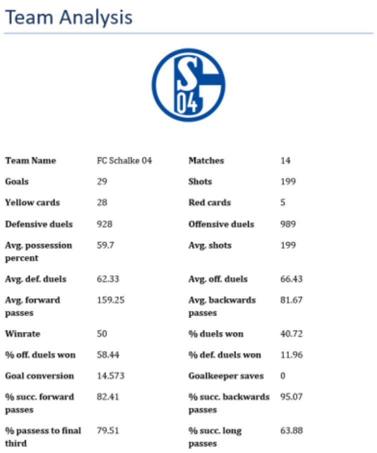
Progetto Finale IDS 2020/21

The report

Requirements

- The report must work with Team B stats, they can come in any format (from a DB, .CSV, .json etc.) a module called “globals.py” must fetch them. By changing this module, we can change the source format.
- The report must be customizable, in the stat-sheet section the analyst can choose which stats to show.
- The advanced stats should not include any math, it must be readable to anyone so only the results should be shown, alongside an explanation.
- The conclusions must focus on the outliers alongside the sample size to know if it is a relevant stat or not.

Stat-sheet



Team Analysis

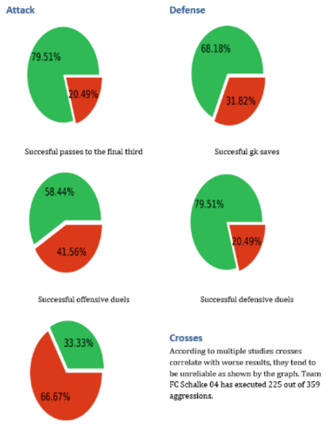
Team Name	FC Schalke 04	Matches	14
Goals	29	Shots	199
Yellow cards	28	Red cards	5
Defensive duels	928	Offensive duels	989
Avg. possession percent	59.7	Avg. shots	199
Avg. def. duels	62.33	Avg. off. duels	66.43
Avg. forward passes	159.25	Avg. backwards passes	81.67
Winrate	50	% duels won	40.72
% off. duels won	58.44	% def. duels won	11.96
Goal conversion	14.573	Goalkeeper saves	0
% succ. forward passes	82.41	% succ. backwards passes	95.07
% passess to final third	79.51	% succ. long passes	63.88

This table displays 24 default values, I have chosen some stats in this example, but it can be customized to display any stat available.

It works independently of what the original raw stats were, the “table_data.py” module fetches the stats from “globals.py”

Its use is mostly as a general overview of team B for a first evaluation and to give better context to the more advanced stats later in the report.

Advanced stats



Attack

Defense

Succesful passes to the final third

Succesful gk saves

Successful offensive duels

Successful defensive duels

Crosses

According to multiple studies crosses correlate with worse results, they tend to be unreliable as shown by the graph. Team FC Schalke 04 has executed 225 out of 359 aggressions.

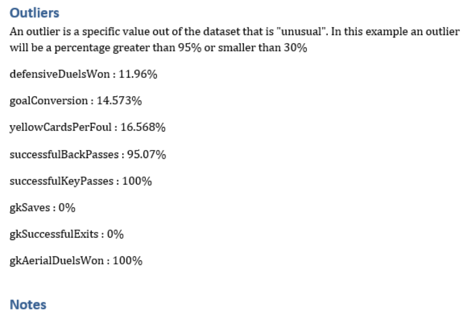
The advanced stats are a different mix of important stats, data visualizations, and graphs. All the calculations are not present on the paper, they are mostly done in the “utils.py” module. Pie charts are a good way to visualize percentages, in this case they are colour coded (green means successful). Any stat can be analysed/visualized here, depending on what the coaching staff wants, this can go on for as many variables as we want.

An import underrated stat I chose for this paper is the number of crosses, as analysed in [2] and [3] this stat tends to correlate with a worse performance, these passes are usually high risk-high reward, and they can lead to losing possession of the ball. Short strategic passes and playing on the sides is a better alternative.

Outliers

A good way for an amateur to start analysing a team is looking for statistical outliers, these are values that are “unusual”. An “unusual” value needs to be defined, choosing the right values is key, especially on a bigger dataset, as selecting a too specific condition may lead to missing an important outlier while selecting a broader range might lead to too many values to analyse.

In this example I chose values smaller than 30% and greater than 95% leading to 8 outputs. These results can lead us to an interesting conclusion, looking at an 11.96% of defensive duels won and a 0% of gk saves a possible logic conclusion is that Team B’s defence is weak.



Outliers

An outlier is a specific value out of the dataset that is "unusual". In this example an outlier will be a percentage greater than 95% or smaller than 30%

defensiveDuelsWon : 11.96%

goalConversion : 14.573%

yellowCardsPerFoul : 16.568%

successfulBackPasses : 95.07%

successfulKeyPasses : 100%

gkSaves : 0%

gkSuccessfulExits : 0%

gkAerialDuelsWon : 100%

Notes

Conclusions

The result of this paper is a prototype of an analysis tool, the final product is a small demo of what it can do, I tried to not go over the 5 pages limit.

The result is just the first step of the full process, in a “real” scenario the analyst would give this output to the coaching staff, to help them with the first assessment (probably working with a lot more stats). The coaching staff would then take their conclusion and write in the notes section what new analysis they want, in our example it could be to focus on Team B’s defence. The analyst can modify the and add any new variable or function related to Team B’s defence to get a second output. This can be repeated any number of times.

The easy, modular structure of the demo and its simple visualization can greatly improve the communication of the team.

References

[1]Nathan Luzum, M. M. (n.d.). The Soccer Analytics Revolution. *sites.duke.edu*.

[2]Geurkink, Y., Boone, J., Verstockt, S., & Bourgois. (2021). Machine Learning-Based Identification of the Strongest Predictive Variables of Winning and Losing in Belgian Professional Soccer. *Appl. Sci.*

[3] Castellano, J., Casamichana, D., & Lago, C. (2012). The Use of Match Statistics that Discriminate Between Successful and UnsuccessfulSoccer Teams. *J. Hum. Kinet*.