Inverting the Pyramid ‘Back’

EAS 503 Project - FIFA World Cup 2018: Inside Out

Archit Singh
Graduate Student - Data Sciences
University at Buffalo
New York, United States
architsi@buffalo.edu

Aditya Sahay
Graduate Student - Data Sciences
University at Buffalo
New York, United States
asahay@buffalo.edu

Shubham Sharma
Graduate Student - Data Sciences
University at Buffalo
New York, United States
ss628@buffalo.edu

***Abstract*—With the insurgence of big data and numerous efficient data mining techniques, a lot of domains have seen a positive delta in their performance KPIs and the field of sports is no exception. Over the years we have seen sports communities like NBA (National Basketball Association) and MLB (Major league Baseball) leverage the potential of growing data to plan their in-match / out-of-match strategies, increase their fan base, strategize the launch of events and what not. Soccer being the most popular sport worldwide has not remained untouched as well. In this project in-match data was used to analyze the performance of players as well as teams by defining metrics that have not been used before. The analysis could help a sports analytics firm or a football management team to look at the performance of their players and the strategies of the opponents which would enable them to take the required action. At the same time, some counter intuitive and hidden insights would be gleaned out, which are not presented to the general audience. The above mentioned goal was achieved by defining metrics that helped identify the strengths and weaknesses of different players and by drawing out insights that could look surprising to people who have been for long exposed to superficial analysis.**

I. Introduction

The project is termed as Inverting the Pyramid ‘Back’ which is inspired by one of the famous books written on football by Jonathan Williams where he talks about how a sudden shift which was non-intuitive resulted in much better playing style and match results. Although we are not literally inverting the pyramid back, we have worked on analyzing the data in such a way which was not done before. We have analyzed the stream data of FIFA World Cup 2018 data to

- Glean out non-intuitive insights
- Present highlights and low-lights of the tournament
- Come up with interesting metrics to determine team’s and player’s performance.

II. Data

The data set was taken from statsbomb.com where they made this data available for free. The following data sets were used:

- Line up
- Events (each event correspond to one activity like pass, shot, goal, etc.)
- Matches (meta-data of the matches played)

The above data was available in JSON format with nested variables. Since MySQL database was preferred for this project, we filtered out the variables of our interest from each of the data sets in the JSON format. We then converted the filtered JSON data and wrote them to CSV files using Node.js. The generated CSV files were used to import the data to MySQL database hosted on Amazon Web Services. We then created additional tables, using this data, to score the performance of various teams and players that were computed using the events data, based on our metrics described later on.

III. Analysis

A. Teams

- We analyzed how the rankings and standings of all the teams changed compared to the pre-world cup period. We analyzed the performance of all the teams in the tournament and compared it to their standings from before the world cup. The post world-cup rankings were decided on how far the teams went in the tournament.
- To help the spectators visualize a teams journey we calculated the points accumulation of different teams after every match. The graph generated shows a step increase in the case of a win/draw. This graph helps us in understanding if the exit of one of the big teams was actually an upset or was it something that was likely to happen.
- We dissected the performance of all the teams to understand their work rate and playing style. To understand the work rate we calculated distance run by the teams for every goal that they scored. This helped us understand some unusual trends which are discussed in the conclusion. On the other side, we also looked at the number of passes made per goal scored by them. This helped us understand their playing style. The combination of these charts told us if the teams were more of smart workers or hard workers. France and Croatia, which were the finalists lie towards the lower end which was counter intuitive for us.
- To enable users look at a particular team and their performance at one place we analyzed the three aspects of the game: Attack, Defense and Midfield. Attack score was given based on Goals scored per game, Difficulty of the shots, number of successful dribbles per game. Defense score was given based on number of blocks, clearances, inverse of goals conceded, and number of

saves made per match. Midfield score was assigned based on Possession, Passing accuracy and number of passes. These three dimensions along with the final performance of the team was shown using a bubble chart with Defense on x axis, Attack on y axis, size of the bubble denoting the Midfield and the color of the bubble representing performance of the team.

B. Players

Players were analyzed by categorizing them based on their playing positions Attack, Midfield and Defense.

- Attackers were analyzed going beyond the regular metrics that are used. Instead of just looking at scoring ratio we analyzed their shooting ability compared to the difficulty of shots they have attempted. This could help us compare two players having same scoring ratio. Bubble chart was used to show the different dimensions.
- Midfielders were analyzed based on both the responsibilities they carry. Their ability to help build the game was assessed using the number of passes made per match. Their ability to defend the defenders was analyzed by looking at the number of times they got dispossessed. Both these metrics were compared by plotting them on a graph and players were segregated in 4 quadrants.
- Analyzing defenders with the help of data is the most challenging task as their role and performance gets significant even when they are off the ball. For example, putting pressure on the opponent, maintaining an offside trap etc. We have tried to cover all the aspects of a defenders game Discipline (number of cards shown), successful and unsuccessful tackles, clearances and blocks. Spider plot was chosen to show these 5 dimensions and to help us compare multiple players at the same time. In figure 1 we compare the performance of Samuel Y. Umtiti with Mats Hummels.
- Since many scouts and teams have a preference to hire players with specific spread on ground during the game, we have plotted heatmaps of different players. This could help teams compare two players with similar results. For instance, Ngolo Kante and Henderson can be seen as similar players when it comes to midfield but by looking at their heatmap we can analyze the width in their games. On comparing the two players who play at the same position but with different playing styles we see that Ronaldo's movement and spread is more towards the goal scoring region (Figure 2) as compared to Hazard who believes in operating up and down the field in order to assist the defense as well as attack (Figure 3).

IV. Results

As mentioned in the goal of the project we were attempting to extract some insights which were counter-intuitive which we were able to achieve with the help of all the analysis done on the highly granular data available. Some of the highlights are:

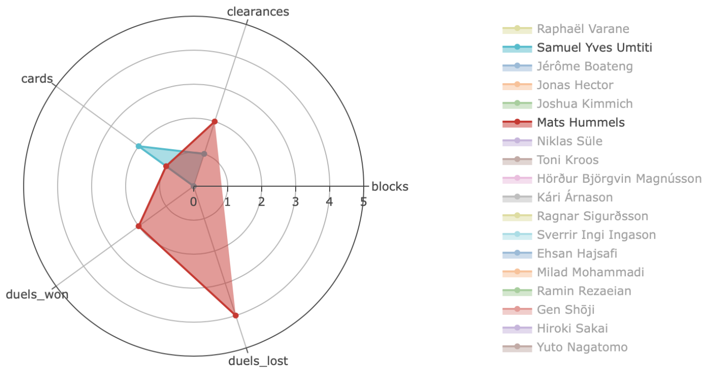

Fig. 1. Spider Plots for Defenders: Samuel Y. Umtiti & Mats Hummels

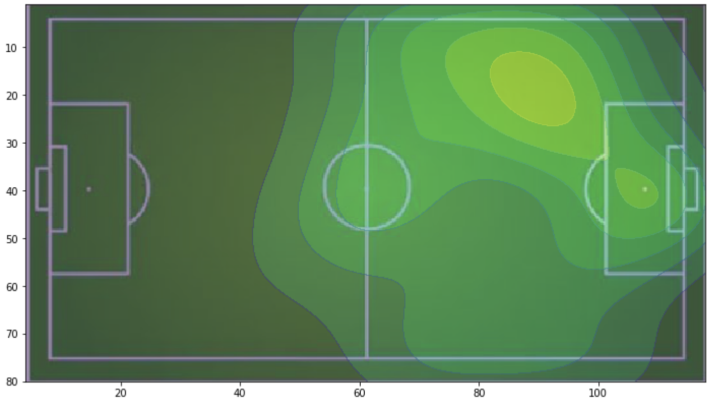

Fig. 2. Heat Map of Christiano Ronaldo

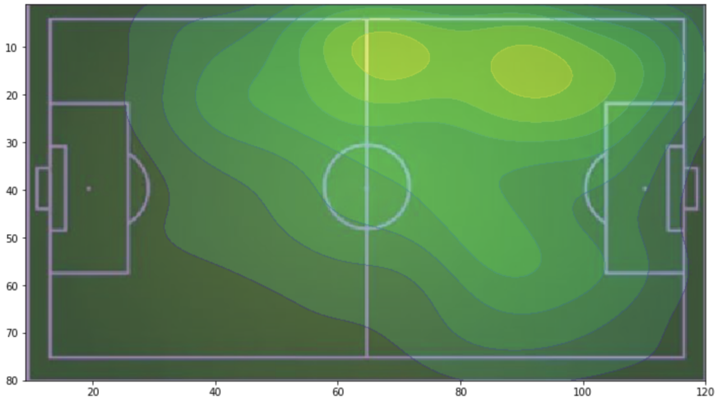

Fig. 3. Heat Map of Eden Hazard

- Although Brazil didnt advance till the end stages of the tournament, Neymar stands out at someone with the best shooting ability as compared to the Golden boot winner who scored 30% of the goals in the form of penalties.
- France stands 13th in the standings when talked about distance run in a match. Also, it stands at 5th in terms of passes made. France has shown us how to excel even with less work rate and less passes.

V. Conclusions

- Although aggregated data make a good cover story of how a match or a tournament went but its the in-match data that helps us unlock the hidden trends.
- When analyzing a players performance, it gets very crucial to look at multiple aspects of their performance which proves helpful in numerous ways.

VI. Future research directions

- We plan to take this forward and incorporate more in match stats. This will all combine to form a dashboard that can be used by teams and clubs.
- We also plan to approach Statsbomb (provider of this data) with a request for data of other world cups in order to see trends across subsequent tournaments.
- We plan to frame an analysis around whether there is a difference in players performance when they play for club and their countries.
- Neural Network models can be used to learn the playing style of the players as well as teams. The model would be used to suggest the best combination of players who can win against a known team.

References

All the data that were used for this analysis, have been taken from StatsBomb [1].

References

[1] Ted Knutson, author, StatsBomb. https://statsbomb.com/2018/09/statsbomb-release-free-fifa-world-cup-data/