The University of Edinburgh

Data Science for Design

Assignment 3 Report

Ioannis Stournaras ID: s1879286
Michail Pourpoulakis ID: s1883345
Pavlos Gogousis ID:s1884197

Data Analysis

As mentioned in our individual reports, our data contain information about the football matches that took place in the world cup qualification stage. Each row represents a different game where the team and 4 different metrics showing the performance of the 11 players appear. These metrics are closeness, betweenness, median X-axis and Y-axis positions. Finally, there is a 'cluster' variable, whose values are between 1-6, that each row is assigned to.

In order to understand how separable the clusters are, we performed Principal Components Analysis (PCA). The analysis helped identifying which clusters are closely related, process that resulted in extracting information from the dataset in an easier manner. Although the data provided are sufficient for an initial analysis of a team, obtaining more data can further separate the clusters. Consequently, the analysis may produce more accurate observations and results. For example, if we get more data describing the outcome of the game, or which team was playing against the examined team when the data were obtained, we might be able to implement a system that predicts the outcome of the game or makes suggestions on how a team, mostly assigned to cluster A, can confront a team assigned to cluster B in order to improve its chances of winning.

Nonetheless, we decided that being able to predict the cluster given the 4 metrics of the 11 players would be an interesting idea, as well as an important feature for football specialists. Using this tool, managers and analysts, would be able to investigate what changes should be made in order for their team to be classified in a different cluster. For this reason, we applied machine learning techniques. First, we split our data in a training set (80% of the data) and a test set (20% of the data) and trained a logistic regression classifier. Then we evaluated our model in the test set and we achieved an accuracy of 94.5%. Given the fact that PCA did not manage to separate all the clusters, due to the fact that highly correlated clusters existed, we achieved an outstanding accuracy, resulting in a very precise model.

Given the fact that our data are suitable for comparisons between teams, players and clusters we decided to implement a web application where interactive graphs are provided to our users. The application is mainly referred to sport analysts and managers since football specific statistics are mainly used for comparisons, which are not so obvious to the general public. An individual could use this information and the analysis conducted in the web application to make suggestions on which players should be called up in order to improve the performance of a team or which formation might be better for a specific team. Moreover, a data set description is provided in case a non-expert user wants to access and use our app. Finally, we ought

to mention that our web-page is not static and it is built on Dash - Plotly.js, React, and Flask.

Our FIFA WORLD CUP 2018 Qualifiers Data Analysis web-page has 6 sections, under which a simple description of the functionality of it is provided.

Home

The home page provides a concise description of our data so the user gets familiar with the statistical metrics in case he is not. Furthermore, a visualisation of our data using PCA is provided so the user gets some intuition on the way clusters are formed and which are separable and which are highly correlated.

Team Betweenness and Closeness

The section Betweenness and Closeness provides to the user the ability to compare two teams based on the mean Betweenness and Closeness of their players' performances. Specifically, the user can select the teams and clusters he wishes to compare and after making his choice, a bar plot will appear comparing the two metrics mentioned before for each of the 11 players.

Team Formation

The tab Team Formation contains both formations, and radar charts. First of all, the user has to choose the 2 teams and select from the available clusters section that he wants to analyse. After each selection, the average formation based on the selected cluster will appear. Then, the user is able to choose from each graph the two players he wants to compare by hovering his cursor over parts of the Graph. The left radar chart will then present the comparison of the two players based on their betweenness, closeness, X-axis and Y-axis positions. Finally, the right radar chart provides a visual comparison of the player chosen from the first team with the average player of the same cluster playing at the same position. The radar chart plot was inspired by the FIFA game developed by EA Sports.

Player Comparison

The Player comparison tab gives the user the opportunity to compare two different players. Players can acquire values between 1 and 11. After selecting the players from the drop-down boxes, the user can also select the statistical metric he is interested in. These metrics provided are the X-axis position, Y-axis position, closeness and betweenness. The bar plot provides

information about the average value of the selected metric for all 6 different clusters.

Builder Formation

Another feature that is provided by our application is to be able to modify a formation and classify the customised system to our set of clusters. To be more specific, the user can choose a specific game from a cluster and a team and the formation will show up. Then, he has the ability to modify the closeness and betweenness values of a selected player, as well as his position on the X and Y axes. After modifying a player, he has to press the 'Submit Player' button. The new position of the player will appear in a second pitch which can be used to compare the initial formation of the team with the custom formation the user decided to use. Once the modification process is finished, users have to press the 'Submit Formation' button and the cluster that the new formation was classified to will appear.

In order to form a viable formation, each player can be assigned values between the minimum and the maximum values observed in the given dataset for each of the 4 metrics. Otherwise, it would be possible to form a team without a goalkeeper, with players playing on only one side of the field or making the striker to be the player from whom the flow of the game is mostly passing by. All these cases are not realistic ones and we can avoid them by limiting the range of the values each player can take.

Betweenness Flow

We have also created a connected graph from which the flow of the game is visible. The thicker black lines that connect players represent a connections between players with extremely high betweenness values. The green medium lines that connect two players demonstrate a good betweenness value for both players, while, lastly, the dashed blue lines mean low to zero betweenness values between players.

We have also differentiated the size of the nodes based on closeness values. Wider-bigger nodes represent a higher closeness value, while smaller ones distinguish lower closeness values. Although one may argue that a lot much information is contained within a single graph, those two metrics are not to be compared independently and make relevant comparisons between players at a single instance.