Analyse statistique de données sportives

Comment conduire le Toulouse Football Club en coupe d'Europe ?

Contexte

Qu'elle porte le nom de Big Data ou d'intelligence artificielle, l'analyse de données est en plein essor dans le domaine du sport. Cette approche a été popularisée notamment depuis 2011 avec la sortie du film *Le Stratège*[1] qui s'inspire d'une histoire vraie d'un analyste statisticien recruté par une club de base-ball. Dans la presse sportive, des articles de plus en plus nombreux font état de l'apport de ce domaine de l'IA dans le management des équipes [1,2,3]. La littérature scientifique est également de plus en plus ouverte à ce type de travaux [4,5,6].

Sujet

Le sujet vise à étudier un jeu de données issu d'une base de données sportives telles que celles disponible sur le site whoscored.com. Ce site compile un grand nombre de paramètres relatifs à des compétitions de football dans le monde entier depuis la saison 1999/2000. Les paramètres sont aussi bien attachés à une équipe (buts marqués, tirs, tackles, interceptions, fautes, cartons...) qu'à un joueur (buts, passes décisives, % passes réussies, fautes commises et subies, hors-jeu...) dans différentes compétitions nationales ou internationales.

Le travail demandé consistera dans un premier temps à définir, avec les encadrants, un périmètre d'analyse (par exemple, les résultats des équipes de la *Premier League* anglaise depuis la saison 2009/2010, les équipes championnes en Europe en 2018/2019...). Ensuite, des analyses exploratoires viseront à mettre en évidence les principales caractéristiques du jeu de données. La suite pourra s'envisager de différentes façons à travers des questions comme : quelles sont les caractéristiques des équipes qui remporte le championnat (*key performance indicator*) ? Ces caractéristiques sont-elles les mêmes dans tous les pays ? Les meilleures équipes (celles qui gagnent les championnats) ont-elles les meilleurs (à définir) joueurs dans leurs rangs ? … Bref, un nombre important de questions pourraient être traitées dans ce contexte... Les méthodes statistiques à mobiliser pour cela sont clairement relatives au *machine learning*, voire à la modélisation et à la sélection de variables. Les développements pourront être réalisés en R ou Python et un soin particulier devra être apporté à la lisibilité des codes en vue d'une ré-utilisation ultérieure.

Références :

[1] Coupe du monde: Comment le Big Data coache l'équipe d'Allemagne. L'express, 2014. www.lexpress.fr/actualite/sport/football/coupe-du-monde-comment-le-big-data-coache-l-equipe-d-allemagne_1553634.html

[2] Big data et football : comment jongler avec les données ? theconversation.com/big-data-et-football-comment-jongler-avec-les-donnees-98477

[3] Sport et Big Data – Quand la science des données donne l'avantage sur le terrain, www.lebigdata.fr/sport-et-big-data

[4] P. Cintia, F. Giannotti, L. Pappalardo, D. Pedreschi and M. Malvaldi, 2015. The harsh rule of the goals: Data-driven performance indicators for football teams, IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, 2015, pp. 1-10.

[5] D. Memmert, D. Raabe, 2018. Data Analytics in Football - Positional Data Collection, Modelling and Analysis, Taylor & Francis Group

[6] H. Janetzko, D. Sacha, M. Stein, T. Schreck, D. A. Keim and O. Deussen, 2014. Feature-driven visual analytics of soccer data, IEEE Conference on Visual Analytics Science and Technology (VAST), Paris, 2014, pp. 13-22.

Contact: Sébastien Déjean (IMT, sebastien.dejean@math.univ-toulouse.fr), Philippe Saint Pierre (IMT, Philippe.Saint-Pierre@math.univ-toulouse.fr), Javier López Sánchez (TUC Tennis et ACC Ramonville, jlsone1@hotmail.com)

1 http://www.allocine.fr/film/fichefilm_gen_cfilm=140005.html