FIFA WC outcomes prediction

Thomas Fernandes

2023-04-12

1. Introduction

1.1 Ce que dit l'article

Cet article examine la prévisibilité des résultats de la Coupe du monde de football en utilisant deux modèles empiriques. Depuis 1993, le système de classement mensuel (couramment appelé coefficient des pays) de la FIFA pour les équipes nationales de football seniors est devenu une source fiable pour juger du potentiel d'une équipe dans la compétition. Cependant, l'article montre que, malgré le nombre de surprises dans les résultats des matchs, les équipes les mieux classées ont généralement de meilleures performances. Les auteurs analysent les résultats des quatre dernières Coupes du monde de la FIFA de 1994 à 2006 et utilisent les classements mensuels de la FIFA pour évaluer les performances des équipes. Deux modèles empiriques sont proposés : un modèle probit et un modèle de régression linéaire MCO. Le premier modèle évalue la probabilité de victoire pour chaque équipe en fonction de son classement FIFA, tandis que le second modèle utilise la différence de buts comme variable dépendante pour voir si les équipes les mieux classées ont généralement des performances supérieures. Les auteurs trouvent des preuves empiriques que, malgré les surprises occasionnelles, les équipes les mieux classées ont généralement de meilleures performances.

1.2 Ce qu'on cherche à faire

A rédiger

2. Traitement des données

Les données liées aux coupes du monde n'étant officiellement plus disponible gratuitement sur le site de la FIFA, il faut se baser sur des données d'utilisateurs qui ont scrap le site de la FIFA dans l'objectif de répliquer ces bases de données.

Classement FIFA : https://www.kaggle.com/datasets/cashncarry/fifaworldranking — Alex Zabrodin Ces données comprennent le classement fifa de 1992 (date de création) à nos jours.

Affluence : https://www.kaggle.com/datasets/abecklas/fifa-world-cup — Andre Becklas Détails des affluences ainsi que des scores des matchs de coupe du monde de 1a première en 1930 à celle de 2014

Coupe du monde 2022 : https://www.kaggle.com/datasets/die9origephit/fifa-world-cup-2022-complete-dataset — Diego Farchione Database complète de toutes les statistiques de chaque match de la coupe du monde 2022

Matchs : https://github.com/jfjelstul/worldcup — Josh Fjelstul Base de données comprenant les scores de tous les matchs de coupe du monde de sa créaion à nos jours

All : https://www.kaggle.com/datasets/piterfm/fifa-football-world-cup — Petro Ivaniuk Base de données complète reprenant les contenant le score de tous les matchs de coupe du monde (1930-2022), les cartons (jaunes et rouges), l'affluence

ALL2: https://www.kaggle.com/datasets/kaito510/fifa-world-cup-match-stats — Kaito G. Contient les données de possession de balle, carton, tir cadrés de 2001 à 2021.

2.1 Traitement du Score et de l'affluence

```
#On importe le premier csv : "FIFA_World_Cup_1558_23 (Attendance)"
WC_score <- read.csv("WC Data/FIFA_World_Cup_1558_23 (Attendance).csv",
                     header = TRUE,
                     sep = ",",
                     stringsAsFactors = FALSE)

#On supprime les colonnes inutiles
WC_score <- WC_score[,c(2,6,7,8,9,10,12)]

#Nouvelle colonne avec la différence de buts
WC_score$Diff <- WC_score$Home.Team.Goals - WC_score$Away.Team.Goals

#On renomme les colonnes
colnames(WC_score) <- c("Year",
                        "City",
                        "Home",
                        "Away",
                        "Home.Goals",
                        "Away.Goals",
                        "Attendance",
                        "Diff")
```

2.2 Traitement du classement FIFA

```
#On importe le deuxième csv : "fifa_ranking-2022-12-22 (Rank)"
FIFA_rank <- read.csv("WC Data/fifa_ranking-2022-12-22 (Rank).csv",
                      header = TRUE,
                      sep = ",",
                      stringsAsFactors = FALSE)

#On garde que les classements au mois de mai de chaque année
FIFA_rank <- subset(FIFA_rank, month(FIFA_rank$rank_date) == 5)

#Maintenant on ne garde que les années de la coupe du monde
FIFA_rank <- subset(FIFA_rank,year(FIFA_rank$rank_date) %in% c(1994, 1998, 2002, 2006,
                                                               2010, 2014, 2018, 2022))
```