Bayesian Sports Analytics

Comparaison modèle statistique vs bookmakers

Paul de Chassey

Problématique

Peut-on prédire les résultats de football aussi bien (voire mieux) que les bookmakers ?

State :

- Résultats très aléatoires
- Peu de buts (événements rares)
- Différences de niveau entre équipes
- Avantage à domicile

Approche

Approche générale

Approche

- Modèle bayésien hiérarchique
- Basé sur les buts marqués
- Implémenté avec Stan

Pourquoi bayésien ?

- Quantifie l’incertitude
- Partage l’information entre équipes
- Interprétation claire des paramètres

Données utilisées

Données

- Source : football-data.co.uk
- Championnat : Premier League
- Saisons : 2019–2020 → 2021–2022

Variables clés

- Équipe domicile / extérieur
- Buts marqués
- Cotes bookmakers (Bet365)

Taille du dataset

- 1140 matchs
- 20 équipes

Données utilisées

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Date	HomeTeam	AwayTeam	HomeGoals	AwayGoals	Result	HomeShots	AwayShots	HomeShotsTarget	AwayShotsTarget	OddsHome	OddsDraw	OddsAway	Seas
2	11/08/2023	Burnley	Man City	0	3	A	6	17	1	8	8.0	5.5	1.33	2023
3	12/08/2023	Arsenal	Nott'm Forest	2	1	H	15	6	7	2	1.18	7.0	15.0	2023
4	12/08/2023	Bournemouth	West Ham	1	1	D	14	16	5	3	2.7	3.4	2.55	2023
5	12/08/2023	Brighton	Luton	4	1	H	27	9	12	3	1.33	5.5	9.0	2023
6	12/08/2023	Everton	Fulham	0	1	A	19	9	9	2	2.2	3.4	3.3	2023
7	12/08/2023	Sheffield United	Crystal Palace	0	1	A	8	24	1	8	3.0	3.3	2.38	2023
8	12/08/2023	Newcastle	Aston Villa	5	1	H	17	16	13	6	1.75	3.75	4.6	2023
9	13/08/2023	Brentford	Tottenham	2	2	D	11	18	6	6	2.75	3.4	2.45	2023
10	13/08/2023	Chelsea	Liverpool	1	1	D	10	13	4	1	2.9	3.4	2.38	2023
11	14/08/2023	Man United	Wolves	1	0	H	15	23	3	6	1.33	5.5	9.0	2023

extrait du dataset *E0_Premier_League.csv*

Modèle hiérarchique

Niveau 1 (matchs)

Modèle statistique

Hypothèse clé

Les buts suivent une loi de Poisson

Goals Poisson(λ)

Pourquoi Poisson ?

- Variables discrètes
- Événements rares
- Standard en modélisation sportive

Niveau 2 (équipes)

Modèle mathématique

Nombre de buts attendus

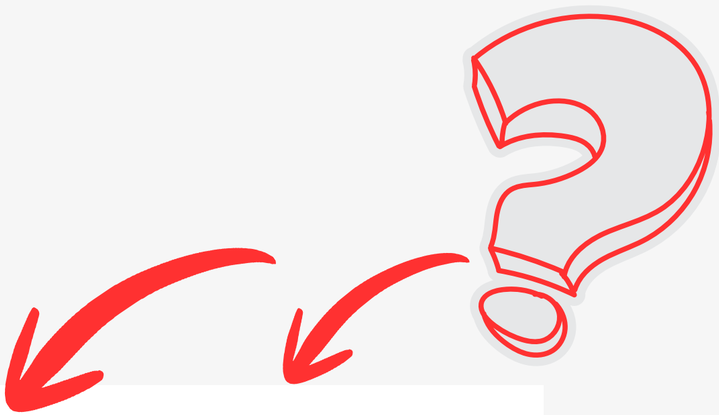
$$\log(\lambda_{\text{home}}) = \mu + home\_adv + attack_{home} - defense_{away}$$

$$\log(\lambda_{\text{away}}) = \mu + attack_{away} - defense_{home}$$

Paramètres :

- μ : niveau moyen du championnat *(niveau global de buts)*
- attack[t] : force offensive de l'équipe
- defense[t] : solidité défensive
- home_adv : avantage à domicile

Modèle mathématique



Problème :

$$\log(\lambda_{\text{home}}) = \mu + home\_adv + attack_{home} - defense_{away}$$

$$\log(\lambda_{\text{away}}) = \mu + attack_{away} - defense_{home}$$

On pourrait dire chaque équipe a un paramètre attaque libre, et un paramètre défense libre, mais :

- trop de paramètres
- les petites équipes avec peu de matchs → estimations instables
- sur-apprentissage

Niveau 3 (championnat)

Modèle hiérarchique

Hiérarchie bayésienne

$$attack_t \sim \mathcal{N}(0, \sigma_{attack})$$

$$defense_t \sim \mathcal{N}(0, \sigma_{defense})$$

Avantages

- Réduction de l’overfitting
- Équipes peu observées stabilisées
- Partage d’information global

et alors ?

Démo !

Merci

Des questions ?