

Projet M1: “Prédire qui gagne le prix pour le meilleur défenseur de la NBA” — comparaison des classements d’attributs

18 mars 2024

1 Évaluer la stabilité des classements

D’abord, on veut savoir si le classement des attributs restent stables entre différentes tranches. Donc pour *chaque méthode* :

- Mutual information
- Pearson’s
- Ridge CV
- Select from Model Logistic
- Select from Model Random Forest

on compare le classement d’une tranche à chaque tranche suivante ainsi que au classement pour toutes les années, ce qui nous donne un tableau de la forme suivante pour chaque méthode :

	1998 – 2002	2003 – 2007	2008 – 2012	2013 – 2017	2018 – 2023	All years
1993–1997						
1998–2002	-					
2003–2007	-	-				
2008–2012	-	-	-			
2013–2017	-	-	-	-		
2018–2023	-	-	-	-	-	

On utilise le Kendall Tau pour comparer les classements.

2 Similarité des classements

Puis, on regarde si les différents classements sont similaires. Donc, pour *chaque tranche* (et toutes les années), on compare les classements de trois méthodes de corrélation entre eux, ainsi que le deux méthodes à partir des modèles :

	Pearson’s	Ridge
Mutual information		
et		
	Random Forest	
Logistic		

3 Davantage de stabilité pour les meilleurs attributs ?

On attend que les attributs qui sont moins importants sont aussi moins stables, on refait donc les exercices précédents pour les top-10, top-20, top-30, top-40, top-50.

4 Enfin, on sélectionne les attributs

Je suppose qu'on va voir que les méthodes restent \pm stables, mais qu'elles sont plutôt différentes entre elles. Par conséquent, on va faire des expériences avec chaque méthode de sélection, en exploitant les valeurs p calculé pour Pearson's. Pour un jeu de training (je vais réfléchir plus sur sa composition et vous envoyer un autre document), on calcule le classement de Pearson's et on sélectionne tous les attributs pour lesquels la valeur $p \leq 0.05$, ça nous donne un nombre k d'attributs. Puis, on sélectionne les k meilleurs attributs selon Mutual information et selon Ridge.

Pour les méthodes "From model", on utilise tous les attributs sélectionnés.