

TP – Introduction à l’Apprentissage Statistique  
ISFA Lyon, Master 2 Actuariat, 2021/2022  
Pierrick Piette

## TP 4 – Boosting

### Librairies nécessaires

xgboost

**Données :** [MedicalPremium.csv](#)

Même base de données qu’au TP 3

### Premier Gradient Boosting

1. Dans quel format les données doivent être mises pour l’utilisation de la librairie xgboost ?
2. Quels sont les hyperparamètres à renseigner dans l’algorithme xgb.train ?
3. Quelles sont les valeurs de ces hyperparamètres pour effectuer un gradient boosting « normal » ?
4. Entrainer un premier GBM avec un taux d’apprentissage de 0.1 et 50 arbres. Analyser l’importance des variables.

### Tuning du nombre d’arbres

1. Quelle fonction permet de tuner le nombre d’arbres optimal ?
2. Faire la cross-validation en testant jusqu’à 200 arbres et mettre en avant l’effet d’overfitting.
3. Quel est le nombre d’arbre optimal ?

### Tuning du taux d’apprentissage

1. On fixe maintenant le taux d’apprentissage à 0.01. Que se passe-t-il lorsque l’on teste avec 200 arbres ?
2. Estimer le nombre d’arbre optimal pour ce taux d’apprentissage.
3. Estimer par cross-validation, le taux d’apprentissage optimal.

### Comparaison avec le CART et la RF

1. Comparer les performances de prédiction du modèle par rapport au CART et RF (cf. TP 3)
2. Qu’en déduire sur les autres paramètres du XGBoost ?