APPLICATION DE L'OPTIMISATION BAYÉSIENNE SUR UNE CLASSIFICATION SALARIALE BINAIRE

Jonathan Moatti, Florent Fettu, Lyes Ould-Ramoul

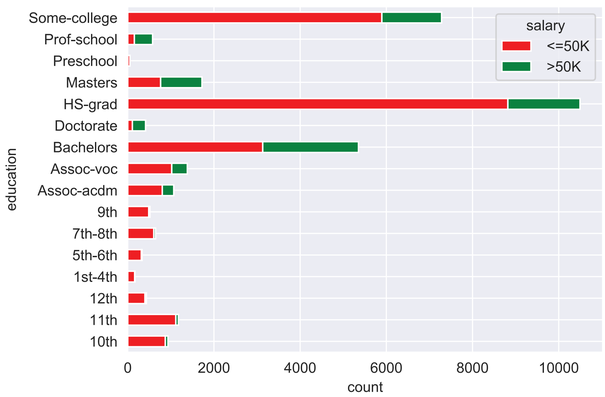
HEC MONTRÉAL

OBJECTIFS

- Tâche de classification pour prédire si un individu gagnera plus de 50k/an
- Optimisation bayésienne des hyper paramètres de 3 modèles d'apprentissage automatique : régression logistique, extreme gradient boosting (xgboost) et réseaux de neurones (keras)

JEU DE DONNÉES

Census Income (UCI) contient 48,842 observations et 15 features provenant de la base de données de recensement des États-Unis en 1994.

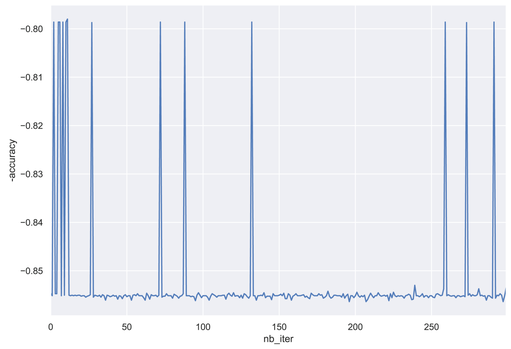


PRÉTRAITEMENT

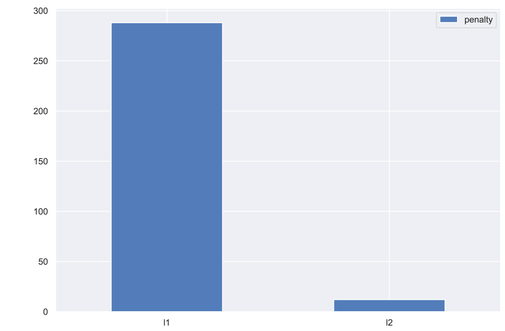
- Traitement des valeurs manquantes
- Transformation des variables catégorielles
- Réduction de la dimensionnalité
- Suppression des doublons

MODÈLES

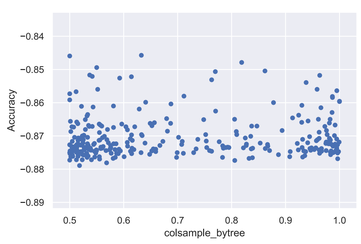
1. RÉGRESSION LOGISTIQUE

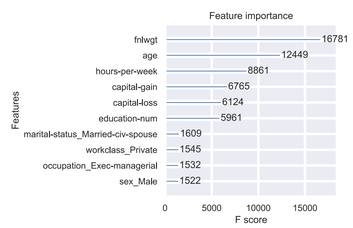


nb_iter	c_space	penalty	accuracy
125	0.767971	l1	-0.855050
126	0.573511	l1	-0.855665
127	0.448232	l1	-0.854743
128	0.177634	l1	-0.854948
129	0.109611	l1	-0.855358
130	0.353192	l1	-0.855665
131	0.705555	l1	-0.855153
132	0.892125	l2	-0.798607
133	0.629477	l1	-0.855153
134	0.951701	l1	-0.855153
135	0.528981	l1	-0.856075



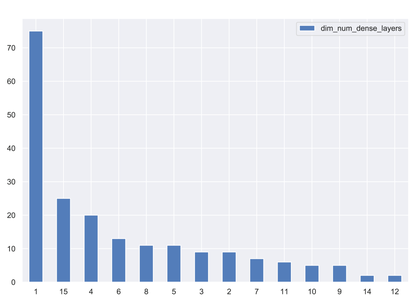
2. EXTREME GRADIENT BOOSTING





3. RÉSEAUX DE NEURONES

dim_learning_rate	dim_num_dense_layers	dim_num_input_nodes	dim_num_dense_nodes	dim_activation	dim_batch_size	Accuracy
0.027749945213301663	4	11	14	sigmoid	338	-0.8572013974189758
1e-05	4	21	2	sigmoid	171	-0.24615857005119324
0.024853382314882457	1	50	46	sigmoid	284	-0.8562794327735901
1e-05	1	1	50	relu	500	-0.4491907358169556



RÉSULTATS

MODÈLES	SCORE EN TEST BENCHMARK	SCORE EN TEST SKOPT	SCORE EN TEST RANDOMIZED
Régression logistique	79,96%	85,38%	84,81%
Extreme gradient boosting	86,92%	87,44%	87,21%
Réseau de neurones	84,60%	85,46%	85,14%

FUTURS TRAVAUX

- Modèle ensembliste : combiner les différents classificateurs par vote ou « stacking » pour améliorer la performance
- « Feature engineering » : créer des nouvelles variables à partir des variables catégorielles et continues

LITTÉRATURES

[1] Navoneel Chakrabarty, Sanket Biswas: « A Statistical Approach to Adult Census Income Level Prediction »

[2] Vidya Chockalingam, Sejal Shah and Ronit Shaw: « Income Classification using Adult Census Data »

[3] Mohammed Topiwalla: « Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting »