Machine learning:
Régression logistique pénalisée et Forêt aléatoire

François Durand Hardy, Rémi Jobard, Quentin Poussier

Rennes 1 - ENSAI

Janvier 2020

Sommaire

Valeurs manquantes

- Problèmes engendrés par les valeurs manquantes:
  -Perte de précisions
  -Biais

- Solutions adoptées
  -Variable correlé pour les variables continues
  -Variable dépendante pour les variables catégorielles
  -Deux modèles
  -Imputation MICE (Cart)

Valeurs manquantes

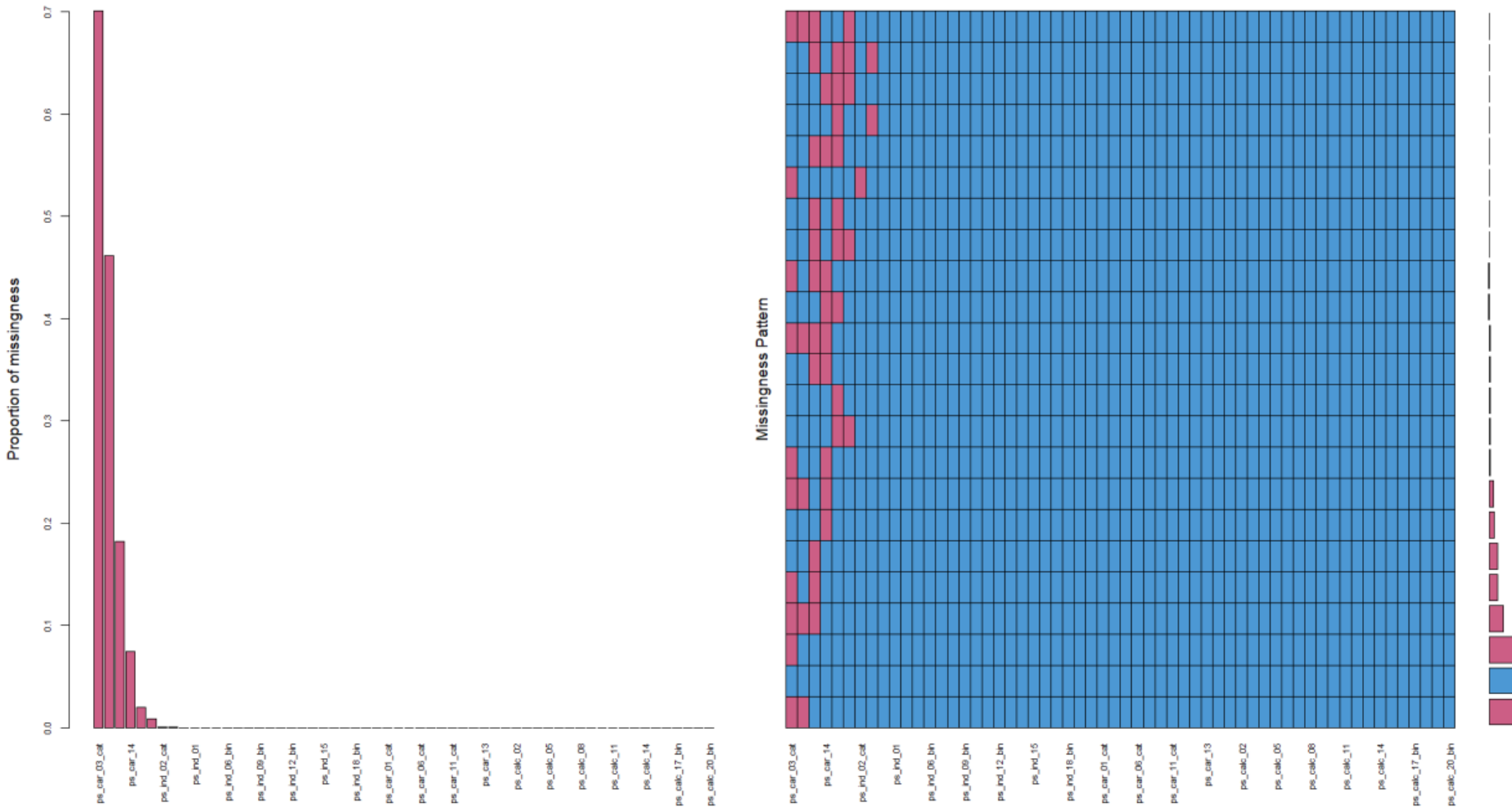


Fig. 1: Représentation des valeurs manquantes

Regroupement des catégories

- Regroupement en respectant les effectifs de la variable cible
- Impact encoding: impact $(m_k) = E[Y|X = m_k] - E[Y]$

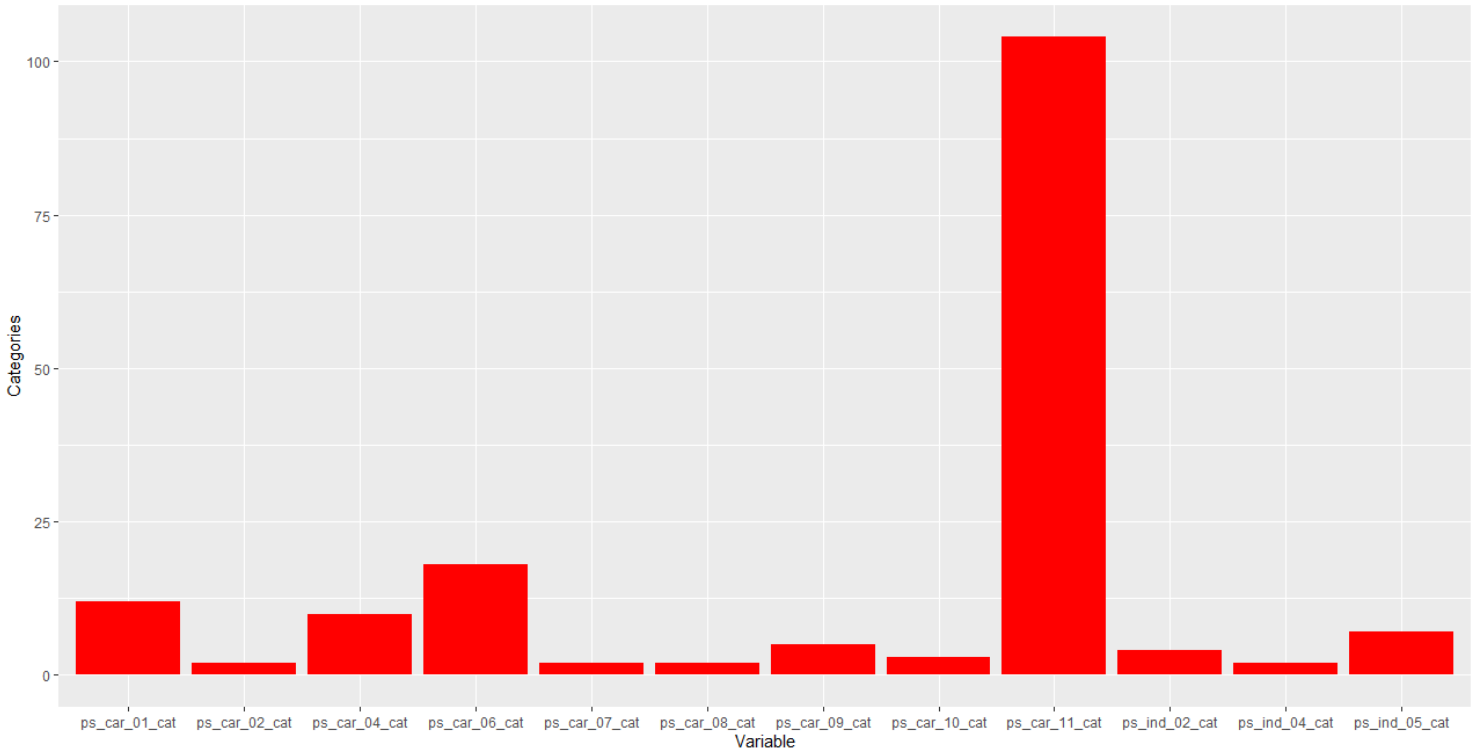
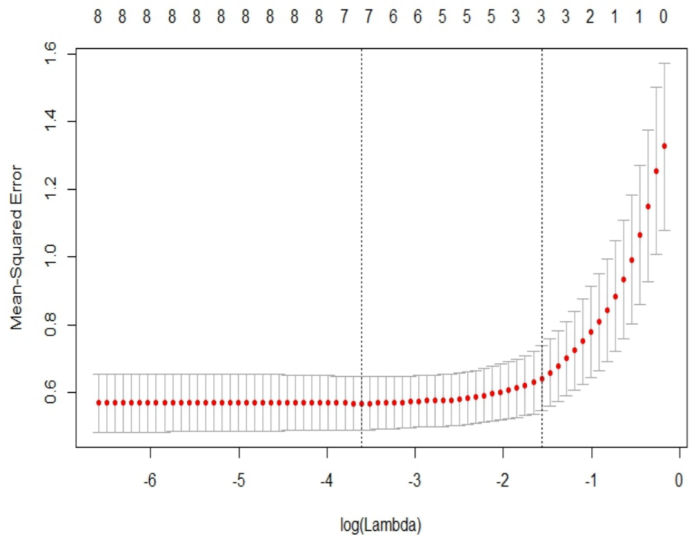


Fig. 2: Nombre de catégories par variable

Régression logistique pénalisée

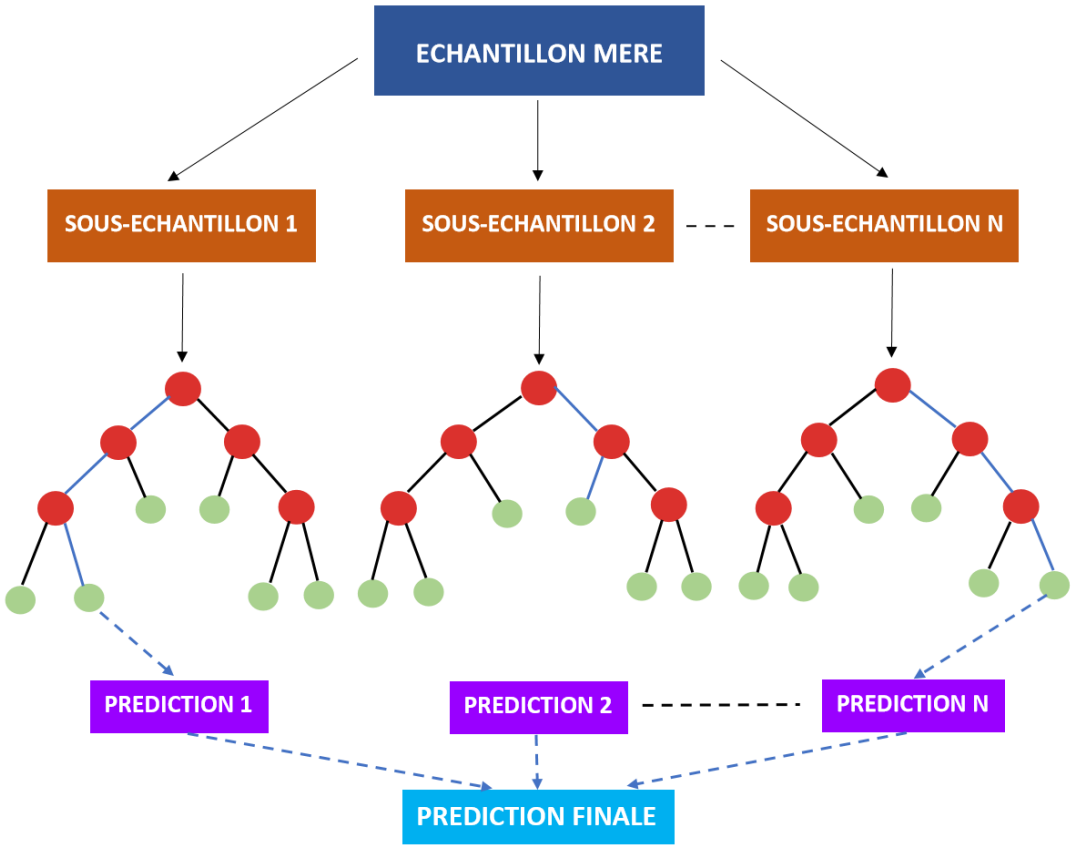
- Trop grand nombre de paramètres augmente la variance car le nombre d'évênements n'est pas nettement supérieur au nombre de variables
- La solution est d'augmenter le biais légèrement pour réduire plus que proportionnellement la variance
- Augmenter la contrainte sur les valeurs que peuvent prendre les paramètres

$$\hat{\beta}_D = argmax(\ell(\beta, D) - \lambda \|\beta\|_1)$$



Forêt aléatoire

- Agrégé différents arbres de décision obtenus à partir de plusieurs sous échantillons bootstrap. Les variables sont également tirées aléatoirement pour répondre au problème d'indépendance des variables.



Sous échantillonnage

- Sous échantillonnage de la classe majoritaire afin d'obtenir une variable cible équilibré.
- Validation croisée: Faire K sous échantillons pour construire les modèles sur chacuns d'entre eux et comparer les K prédictions. Choisir les paramètres (λ ou F le nombre de variables pris dans chaques arbre)

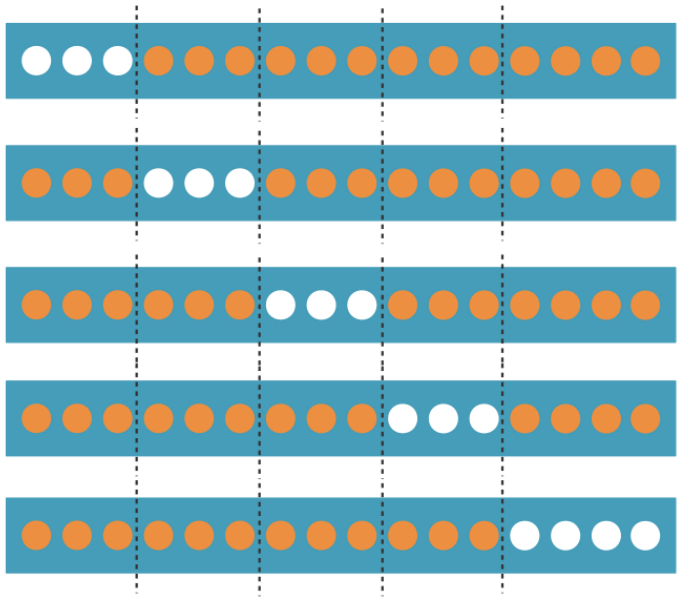


Fig. 5: Validation croisée

Optimisation des modèles

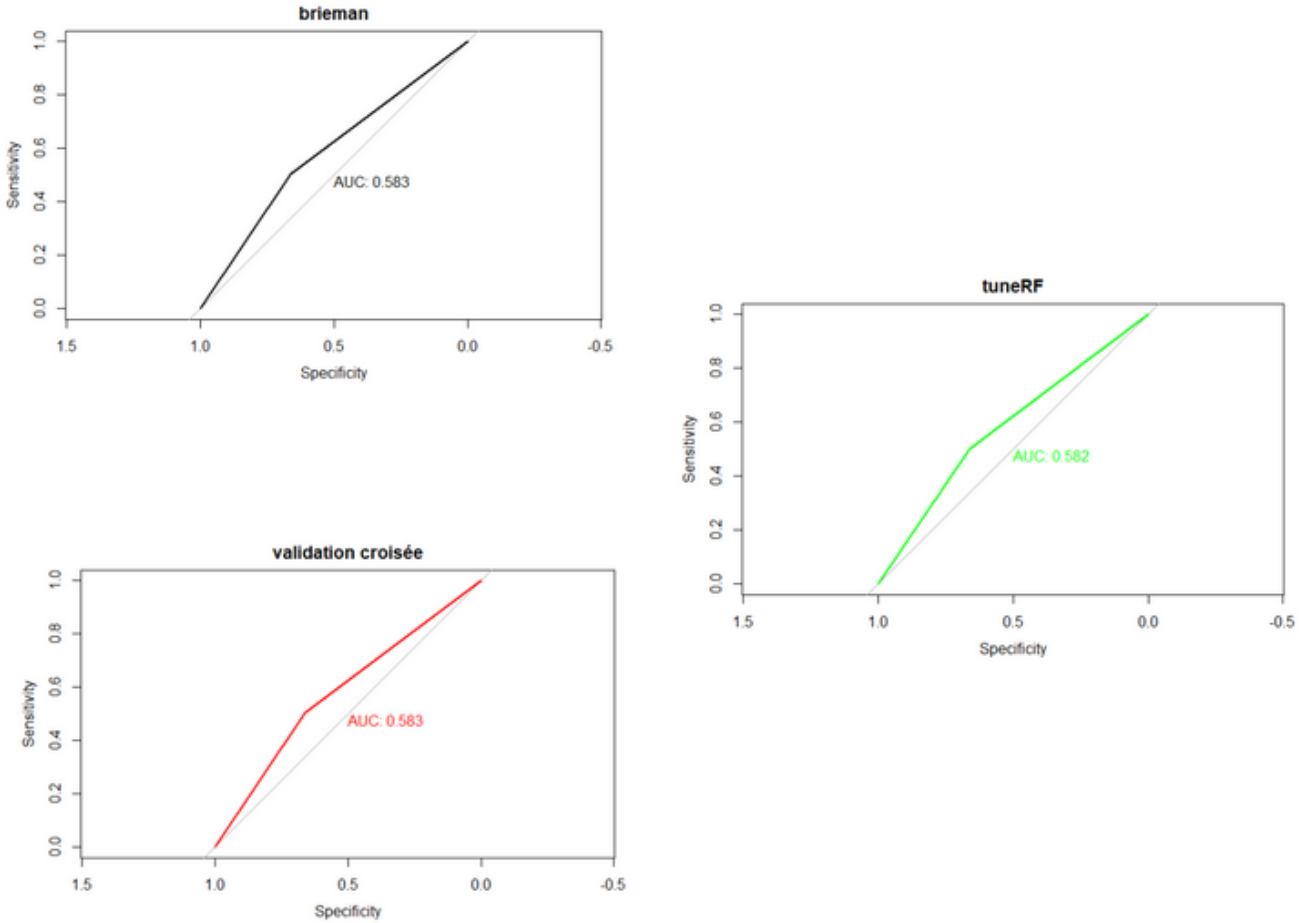
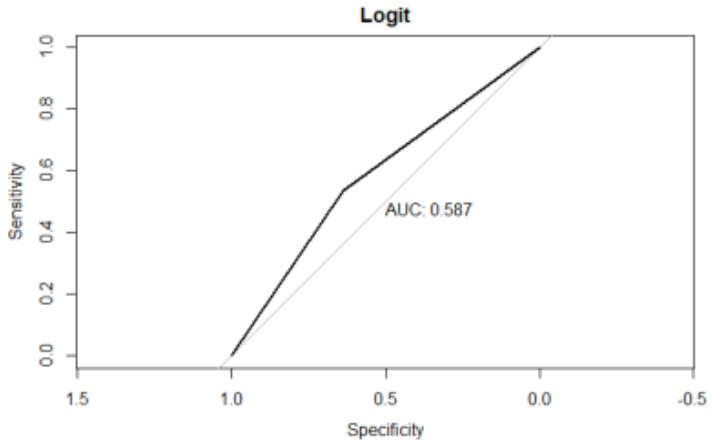
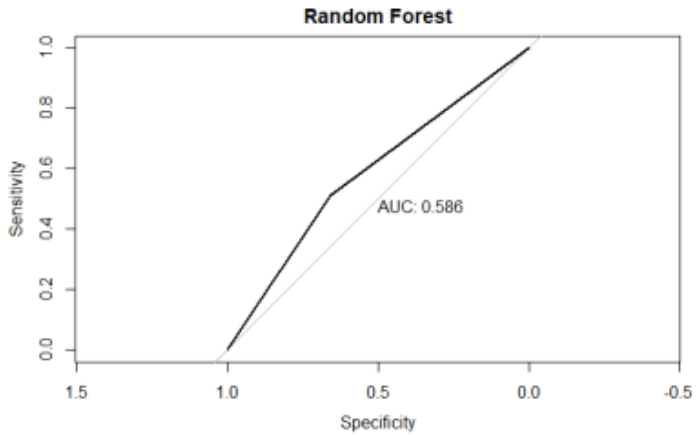


Fig. 6: Courbe ROC des modèles de forêts aléatoires

Généralisation des modèles





	Logit pénalisé	Random forest
Indicateur de Rappel	0,54	0,51

Fig. 7: Courbe ROC et indicateurs de rappel

Bibliographie

- Breiman L, 2001 *Random Forests*, Statistics Department, University of California.
- Donzé L, 1999 *L'imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse es données : l'enquête 1999 sur l'innovation*, Ecole polytechnique fédérale de Zurich.
- Drummond C et Holte R, *Class Imbalance, an Cost sensistivity: why Under-Sampling beats Over-sampling*, National research Council Canada.
- Elhassan AT et al., 2017 *Classification of imbalance Data using Tomek Link (T-link) combined with random under-sampling (RUS) as a Data Reduction Method*. Global Journal of Technology and Optimization
- Gower J, 1986 *Metric and Euclidean properties of dissimilarity coefficients*, Journal of Classification.
- Imbert A, Vialaneix N, 2018 *Decrire, prend en compte, imputer et évaluer les valeurs manquantes dans les étuddes statistiques : une revue des approches existantes*, Journal de la Société Française de Statistique Vol. 159 No. 2.

Bibliographie

- Li C, 2014*Little's test of missing completely at random, Northwestern University.*
- *Marta Avalos. Sélection de variables avec lasso dans la régression logistique conditionnelle*. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009.
- Palm R et IEMMA A. F, 1995 *Quelques alternatives à la régression classique dans le cas de la colinéarité*, Revue de statistique appliquée, tome 43, n2 p5-33.
- Saito T, Rehmsmeier M (2015) *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*.
- Tomek, I 1996, *An experiment with the edited nearest neighbor rule*. IEEE Transactions on sytems, man and cybernetics.
- Van Buuren S et al., 2011 *Multivariate Imputation by Chained Equations in R*, Journal of Statistical Software Vol 45.
- Wilson, D.R. Martinez, T.R. *Machine Learning* (2000)
- Zumel N et Mount T, 2019 *Processor for predictive modeling*.