# Modèle de prédiction de risque aux assurances

**Rémi JOBARD[1], François DURAND [2] et Quentin POUSSIER[3]**

[1,2,3] Master 2 EDP, Université de Rennes 1, Rennes, France.

E-mail : [1]remi.jobard@etudiant.univ-rennes1.fr
[2]francois.durand-hardy@etudiant.univ-rennes1.fr
[3]quentin.poussier@etudiant.univ-rennes1.fr

## Abstract

Le choix d'un modèle de machine learning n'est pas chose aisé. Nous développons à cet effet une méthodologie décrivant la démarche à suivre pour appliquer deux méthodes de machine learning : La régression logistique pénalisée et la forêt aléatoire. Ces méthodes d'estimations nécessitent un travail préliminaire pour le traitement des données afin de maximiser la précision potentiellement dégradée par les valeurs manquantes, mais également de diminuer les potentiels biais dû à la présence de colinéarité. Ces modèles sont ensuite testés et validés grâce à la courbe ROC et aux indicateurs de rappel. Leur construction se fait sur différents sous échantillons permettant de construire et de tester le modèle sur des jeux de données différents. Cette méthodologie s'appuie sur un exemple concret d'estimation du risque de réclamation d'un assuré pour l'année suivante.

Mots clés : Risque, Forêt aléatoire, Régression logistique pénalisée.

---

## Introduction

Qu'il s'agisse de méthode d'apprentissage simple ou avancée, les algorithmes de machine learning demandent un travail préliminaire sur les données. En effet ces algorithmes sont souvent utilisés sur de vastes données ce qui implique, une diversité de données et des valeurs manquantes. Cependant cela donne aussi une opportunité de construire des modèles sur un échantillon différents de celui servant à les tester, du fait de la taille de la base. Pour optimiser les estimations, le data analyste se doit donc de traiter au mieux les données avant d'appliquer un modèle.

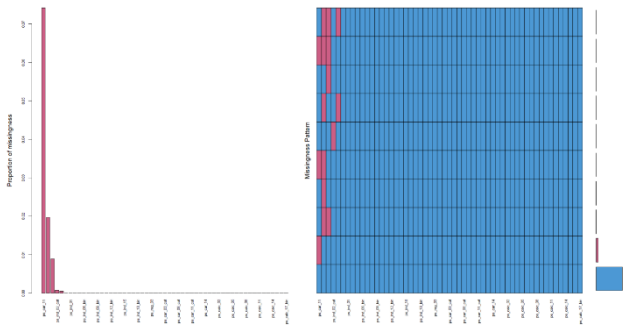
Cet article méthodologique s'appuie sur une application pour illustrer sa démarche : Dans le cadre d'un travail pour une société d'assurance, on cherche à construire un modèle qui prédit pour chaque assuré sa probabilité de déposer une réclamation au cours de la prochaine année. Pour ce faire, deux modèles de machine learning sont utilisés : D'une part, la régression logistique pénalisée qui permet d'estimer et de réduire le nombre de paramètres d'un modèle en rajoutant un critère de pénalisation lors de la maximisation de la vraisemblance. D'autre part, la méthode random forest qui agrège un ensemble d'arbre obtenus par différents échantillons bootstrap et dont les variables ont aussi été tirées aléatoirement. La démarche aura pour but d'optimiser au mieux l'utilisation de ces modèles. Une première partie sera donc consacrée au traitement des données. Les valeurs manquantes et les variables catégorielles font donc l'objet d'une analyse approfondie. Un second temps servira à décrire les aspects théoriques des modèles, les techniques de validation et de comparaison entre ces derniers. Enfin, la dernière partie sera axée sur l'application concrètes des modèles et l'analyse des résultats.

## 1. Traitement des données

Tout d'abord il est primordial d'analyser les valeurs manquantes. Ces valeurs posent deux principales difficultés : une perte de précision et d'importants biais dans les modèles. Pour répondre à ces problèmes, deux stratégies ont été utilisés.

La première concerne les variables avec un fort taux de valeurs manquantes. L'information contenu dans ces variables est substituée par des variables leur étant très corrélé (s'il s'agit d'une variable quantitative) ou très dépendante (s'il s'agit d'une variable qualitative). Cela signifie qui si une variable A est fortement corrélé à une variable B. Alors, la variable B peut remplacer les informations contenues dans la variable A qui est alors supprimé.

Ensuite, pour les variables restantes, une imputation est réalisée. Les valeurs manquantes sont remplacées par des valeurs estimées. L'imputation peut se faire de différentes façons[1] (Alyssa Imbert et Nathalie Vialaneix 2018). Dans l'application, la méthode MICE (van Buuren and Oudshoorn 2000) (Multivariate imputationby chained equations) a été retenue. Elle fait partie des méthodes qui ne tiennent pas compte de la distribution jointe des variables ce qui permet de l'appliquer simplement. En outre, c'est l'algorithme CART (Classification And Regression Trees par Lisa Doove et al. 2012) qui est utiliser. La nature des valeurs manquantes à déterminer le choix du modèle d'imputation. Ici, les données sont manquantes de manière aléatoire. En effet on remarque sur le graphique suivant que les valeurs manquantes en semblent pas avoir de structure particulière. L'imputation par arbre est alors appropriée puisqu'elle s'applique à ce type de données.



Certaines variables contiennent un très grand nombre de catégories ce qui accroit naturellement le nombre de paramètres. Sans regroupement, les catégories avec un faible effectif auront donc mécaniquement un coefficient plus faible. Or, cela amènerait à les exclure du modèle les petites catégories presque systématiquement. Pour remédier à ce problème les catégories sont regroupées pour obtenir au moins 5% d'effectifs dans chaque catégorie. Nous conservons alors une information moins susceptible d'être exclue d'emblée.

Ainsi les variables ps_ind_02_cat, ps_ind_05_cat, ps_car_04_cat, ps_car_01_cat, ps_car_09_cat se sont vues modifiées pour atteindre les 5 % d'effectifs minimum dans chaque classe. Ces regroupements ont été réalisés conditionnellement à la variable d'intérêt. Les répartitions jointes sont respectées.

Une des variables comprenait à elle seule 104 catégories. La stratégie de regroupement appliquée ici est celle de l'impact encoding (Zumel 2012). Elle consiste à regrouper les catégories d'une variable par impact sur la variable d'intérêt. Cette méthode est donc d'une grande utilité lorsqu'une variable compte un nombre de catégorie trop important pour être simplement regroupée par effectifs. Il s'agit donc ici d'un regroupement par impact alors que le regroupement précédent se fait uniquement en respectant la répartition des effectifs.

Une distribution asymétrique de la variable target a été observée : il y a une surreprésentation de non réclamation (96%) par rapport aux individus qui ont débuté une procédure de réclamation (4%). Cela pose un problème car les modèles présentés dans cet article (régression logistique pénalisée et forêt aléatoire) dépendent fortement des proportions des observations, il devient alors nécessaire de procéder à un rééquilibrage. Deux options sont possibles : le sous-échantillonnage de la classe majoritaire (under sampling) ou le sur-échantillonnage de la classe minoritaire (over sampling). Le sur-échantillonnage a été rapidement écarté car il y a énormément d'observation dans la base de données et le déséquilibre est très important, la duplication d'individu engendrerait alors inévitablement du sur-apprentissage. De plus, il est toujours plus simple d'enlever des données plutôt que d'en créer des factices. Cependant, le sous-échantillonnage aléatoire simple peut provoquer une perte d'information dans la mesure où des observations importantes peuvent être écartées. Pour veiller à ce problème, il faudrait contrôler les individus qui feront partis de l'échantillon grâce aux méthodes Edited Nearest-Neighbor (Wilson 1972) , Near Miss ou Tomek Link (Tomek 1976) pour ne citer qu'elles. La méthode Tomek Links étant la plus communément utilisée. En deux mots elle consiste à détecter les individus de la classe majoritaire qui sont proches des individus de la classe minoritaire (appelés « points frontières ») et de les écarter lors du sous-échantillonnage de façon à ce qu'il ne reste que des individus éligibles.

1 Il y a notamment les imputations simples qui consistent à estimer les valeurs manquantes. De leurs côtés, les imputations multiples créées plusieurs valeurs possibles. La valeur permettant la meilleure estimation est ensuite retenue.

Nous voulions appliquer cette méthode en utilisant comme distance HEOM qui permettait de calculer des distances sur des données mixtes. Toutefois, du fait du nombre trop important d'observation dans la base il n'est pas possible de calculer une distance entre chaque individu : cela ferait 416000*416000 opérations et le temps de calcul serait trop élevé. Nous n'avons pas trouvé de méthode de sous-échantillonnage qui permettait de prendre en compte cette problématique. Nous sommes donc restés sur l'idée de sous-échantillon aléatoire simple : toutes les observations de la classe minoritaire sont conservées et le même nombre d'observations sont sélectionnées aléatoirement dans la classe majoritaire afin d'obtenir une distribution 50-50.

## 2. Modélisation de la probabilité de réclamation

Deux types de modèle sont utilisés pour expliquer la variable cible, le premier est un modèle de régression logistique pénalisée et le deuxième est un modèle de forêt aléatoire.

Un nombre important de paramètres peut entrainer un phénomène de multi colinéarité. On parle de multicolinéarité lorsqu'une variable explicative est une combinaison linéaire d'une ou plusieurs autres variables. Ce phénomène entraine une augmentation de la variance des paramètres et peut conduire à la non significativité de certains. Notre application comptant 59 variables, le modèle logistique simple n'est donc plus adéquat. La régression logistique pénalisée Lasso est alors très adaptée à ce problème. Elle permet en effet de retenir un sous ensemble de variables en annulant certains coefficients suivant un seuil défini préalablement. Le principe est d'ajouter une contrainte supplémentaire pour la maximisation de la vraissemblance. Cette contrainte réduit les plages que peuvent prendre les paramètres estimés. Ainsi, on limite les coefficients avec trop de variation. Il existe également une variante nommée Ridge. Cette dernière consiste, non pas à annuler des coefficients, mais à les atténués. La régression logistique pénalisée Lasso est donc préférable dans notre cas puisqu'elle permet une sélection des variables.

La régression logistique pénalisée Lasso repose sur un critère : le Lambda. Ce critère et en fait une contrainte supplémentaire dans la résolution du maximum de vraisemblance. Il s'introduit comme un terme négatif dans la fonction de maximisation. Ainsi, plus le lambda est proche de 0, plus on se rapproche d'une régression logistique classique. En revanche, lorsque le lambda tend vers l'infini, le critère de pénalisation sera plus important et de moins en moins de paramètres seront retenus.

Pour déterminer le lambda optimal, on retient celui minimisant l'erreur de prédiction. Pour ce faire, il suffisait de procéder par validation croisée. La validation croisée est utilisée très fréquemment comme stratégie de validation. Elle consiste à séparer un échantillon en K groupes qui seront utilisés alternativement soit pour élaborer un modèle, soit pour le tester. Pour chacune des estimations il y a une prédiction (sur les groupes utilisés pour tester) et il est possible d'établir un taux de bon classement pour chaque modèle ce qui en fera au total K. Après avoir représenter les différents lambda en fonction de l'erreur de prédiction, le plus grand lambda qui minimise l'erreur de prédiction est celui retenu pour le modèle.

La même méthode (validation croisée) a été utilisée pour déterminer les paramètres optimaux dans le modèle forêt aléatoire. Développé par Brieman et al. en 2001, ce modèle permet de réaliser des arbres de classification sur des sous échantillons bootstrap tirés aléatoirement et avec remise. L'idée est d'ensuite agréger tous ces arbres ensemble. L'intérêt de cette méthode est de réduire considérablement la variance de l'estimateur qui est le principal défaut des arbres de classification classiques sans en modifier le biais. L'inconvénient est que l'hypothèse principale de cette méthode qui est l'indépendance des variables qui sont dans chaque arbres n'est pas validée. Il faut donc introduire une nouvelle source d'aléa en sélectionnant aléatoirement des variables dans chaque arbre constituant la forêt.

Le nombre de variable tiré est crucial car si trop de variables sont sélectionnées, cela va entrainer une augmentation de la corrélation entre chaque régresseur (ici des arbres) ce qui a pour conséquence de réduire la précision de l'estimateur agrégé. Trop peu de variables choisies entrainerait une augmentation du biais des arbres. En général, $\sqrt{K}$ variables sont sélectionnées avec K le nombre total de variable. Afin de choisir le nombre de variable optimal, ce paramètre a été déterminé par 3 méthodes différentes. Comme on l'a indiqué précédemment, la première méthode est par validation croisée. Il sera de cette façon plus adapté à notre jeu de donnée et la validation croisée réduit fortement le risque de sur-apprentissage. La seconde méthode utilise les erreurs Out-Of-Bag introduites par Brieman. Pour chaque observation de l'échantillon d'apprentissage, l'ensemble des arbres de la forêt qui ne contiennent pas l'observation seront utilisés pour prédire cette observation. De cette manière, il est possible d'estimer une erreur à partir de cette prédiction et de la valeur réelle de l'observation. Enfin, la dernière méthode est celle prescrite par Brieman dans son article paru en 2001 qui est de prendre int(log(2*M + 1)) si la majorité des variables sont qualitatives.

## 3. Validations du modèle

Pour obtenir la meilleure validation de nos modèles et optimiser la généralisation de ceux-ci, les données sont séparées en 2 bases. La première servira à la construction des modèles et au choix du meilleur modèle entre chaque algorithme. Pour rappel, cette base ne sera pas utilisée entièrement en raison du sous-échantillonnage. Cet échantillon de la base d'apprentissage sera lui-même séparé en deux, la première partie servira à l'apprentissage des modèles (90% des données) et la seconde sera utilisée pour évaluer les modèles et les comparer entre eux (10% des données). Enfin, la deuxième base appelée « test » servira à mesurer la généralisation des modèles sur un grand jeu de donnée qui ne sera pas équilibrée préalablement.

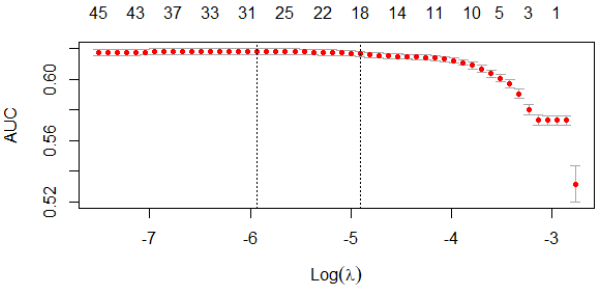
Pour comparer la performance des modèles sur le sous-échantillon test de la base d'apprentissage, c'est la courbe ROC démocratisé par Swets et al (2000) qui sera utilisée. Cette méthode utilise la matrice de confusion mettant en opposition les données réels et les données prédites Les vrais positifs sont les valeurs positives qui ont été prédites comme positives pas le modèle. A l'inverse, les faux négatifs sont des valeurs prédites comme négatives alors qu'elles sont dans la réalité positives. Dans cette matrice, la sensibilité (taux de positifs correctement prédit) sera tracée en fonction de la spécificité (taux de négatifs mal prédits). Le but est suite au tracé de cette courbe de maximiser l'air sous celle-ci (nommée AUC pour Area Under the Curve).

Pour tester la généralisation des modèles, la courbe ROC n'est plus adapté car les données ne sont pas déséquilibrées. L'indice AUC n'étant plus exploitable, c'est l'indice de rappel qui sera préféré. Cet indicateur est aussi déterminé par le biais de la matrice de confusion. Aussi appelé Taux de vrai positif il se calcule : VP/(VP + FN) avec VP les vrais positifs et FN les faux négatifs. Il est privilégié lorsque nous cherchons à détecter la part d'individu correctement détectés parmi l'ensemble des individus pertinents. Comme nous sommes dans le cas d'un risque il est plus pertinent de prendre plus d'individus au risque de se tromper plutôt que d'en écarter un certain nombre.

## 4. Réalisation des modèles

Dans cette partie nous représentons les résultats par des courbes ROC. Cette courbe trace le taux de vrais positifs par rapport au taux de faux positifs. Ainsi, les performances du modèle sont représentées sous forme d'une courbe. L'aire sous la courbe nous donne une indication sur la qualité du modèle à estimer correctement un évènement. Si l'aire sous la courbe vaut 1 cela signifie que le modèle pourra déterminer à 100% si la variable cible vaut 1 ou 0.

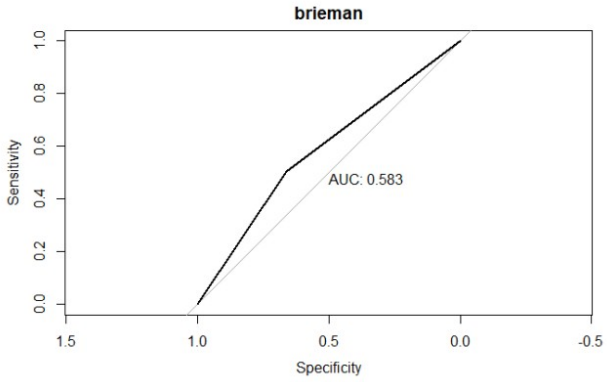
Comme expliqué précédemment, le choix du lambda dans le logit pénalisé a été déterminé par la validation croisée. En s'attardant sur le graphique suivant, on observe les différents logs de lambdas et les AUC qui leur sont associés.



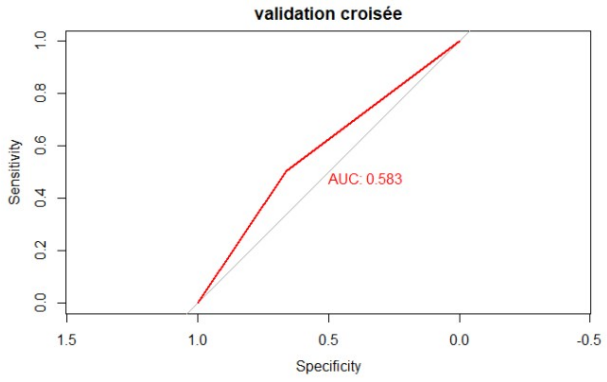

Le tracé vertical placé à gauche correspond au lambda min et celui de droite au lambda les. Le premier correspond au lambda maximisant L'AUC et le deuxième la plus grande valeur possible du lambda avec une erreur à plus un écart type de l'erreur de lambda min. Dans le but de maximiser la parcimonie du modèle, c'est le lambda lse qui est retenu.

L'optimisation du nombre de variables choisies dans notre random forest par les trois méthodes d'optimisation nous donnent les AUC suivantes sur l'échantillon train :

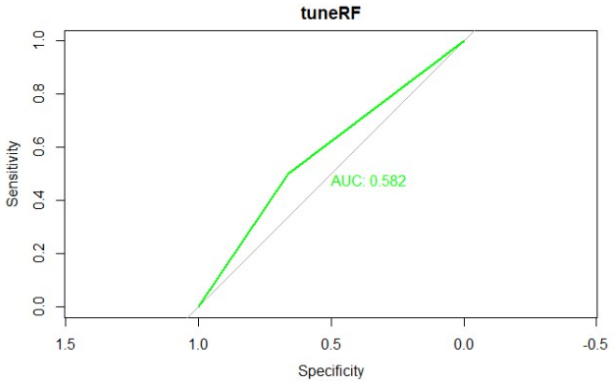
La méthode de Brieman :




La validation croisée :



Le TuneRF :




En comparant les AUC, on détermine donc que la méthode de sélection optimale du nombre de variable est la méthode de Brieman.

Après avoir défini le modèle qui sera utilisé comme comparaison avec le logit pénalisé, nous allons les comparer sur leur taux de rappels.

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

Dans le tableau ci-dessus, on peut voir que le modèle avec le taux de rappel le plus élevé est le logit pénalisé. Ce modèle est donc le modèle qui se généralisera le mieux sur d'autre données.

Une critique peut être cependant apportée à ses résultats. En effet on peut voir que les AUC des deux modèles restent faible et que le pouvoir prédictif en est ainsi réduit.

Truffery S (2007) énonce avec détails les raisons pouvant expliquer la performance d'un modèle et une hypothèse peut être faite pour expliquer la performance de nos modèles. En effet il est dit que la discrétisation des données continues pouvait amener un apport de performance du modèle.

## Conclusion

Les modèles de machine learning demande une préparation des données qui n'est pas négligeable. Elle conditionnera fortement les résultats des modèles. Beaucoup d'algorithme sont, en effet, basées sur de nombreuses itérations ou les données vont être utilisées plusieurs fois. La structure de ces dernières à donc un poids très important. Après avoir rendu les données utilisables, il est fondamental d'analyser les variables pour choisir les modèles et les techniques de sous échantillonnage. Enfin, le choix des indicateurs de validation et de comparaison des modèles est tout aussi important. Comme nous l'avons vue précédemment, les indicateurs peuvent être trompeur ou source d'illusion. Identifier au préalable ces potentiels biais permet de produire des modèles d'une meilleure qualité. Il serait donc intéressant d'effectuer une application en comparant un panel de méthodes différentes et de comparer les résultats pour réussir à quantifier les divergences qu'ils peuvent engendrer.

Pour ce qui est du choix du modèle optimisant la généralisation, il s'avère que le choix d'un modèle logit pénalisé. Cependant la qualité de prédiction reste faible reste faible dans les deux cas.

## References

[1] Breiman L, 2001 *Random Forests,* Statistics Department, University of California.
[2] 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.
[3] Drummond C et Holte R, *Class Imbalance, an Cost sensistivity: why Under-Sampling beats Over-sampling,* National research Council Canada.
[4] 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
[5] Gower J, 1986 *Metric and Euclidean properties of dissimilarity coefficients,* Journal of Classification.
[6] 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 Voo. 159 No. 2.
**[7]** Li C, 2014 *Little's test of missing completely at random,* Northwestern University.
[8] Marta Avalos. *Sélection de variables avec lasso dans la régression logistique conditionnelle*. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009.
[9] 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, n°2 p5-33.
[10] Saito T, Rehmsmeier M (2015) *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*.
[11] Swets, J.A., Dawes, R.M., Monahan, J., 2000. *Better decisions through science*. Scientific American 283, 82–87.
[12] Tomek, I 1996, *An experiment with the edited nearest neighbor rule.* IEEE Transactions on sytems, man and cybernetics.
[13] Truffery S,2007, *Améliorer les performances d'un modèle prédictif: perspectives et réalité,* Data Mining et Apprentissage Statistique, vol. RNTI-A-1, pp.45-72.
**[14]** Van Buuren S et al., 2011 *Multivariate Imputation by Chained Equations in R,* Journal of Statistical Software Vol 45.
[15] Wilson, D.R. & Martinez, T.R. Machine Learning (2000)
[16] Zumel N et Mount T, 2019 *Processor for predictive modeling.*