Adaboost for Neural networks applied to an insurance Database

Lucien Ledune

16 april 2019

Abstract

This will be the abstract of the document

1 Introduction

Partout dans le monde, la facilité montante de l'accès à une puissance de calcul importante a motivé un grand nombre de changement dans la façon d'approcher les problèmes, et ce dans la plupart des domaines de recherche. Cette avancée technologique a permi d'utiliser des techniques de modélisations et de prédictions qui étaient jusqu'à lors trop longues à mettre en place pour avoir un réel intérêt pratique. Le secteur des assurances n'a pas été épargné par le phénomène et évolue avec son temps. Le machine learning[1], outil de modélisation très puissant, est maintenant facilement accessible et les compagnies d'assurances souhaient en tirer le meilleur parti en matière de prédiction et d'aide à la décision.
La modélisation de prédiction n'est pas nouvelle dans ce secteur, mais les algorithmes utilisés changent avec l'essort actuel de la technologie. Ainsi nous pouvons maintenant facilement utiliser des réseaux de neurones afin de répondre aux problèmes pour lesquels des algorithmes moins efficaces étaient utilisés. Le but de ce travail est de montrer que ces nouveaux[2] algorithmes sont efficaces et représentent un espoir d'amélioration pour le futur. TODO : expliquer le mémoire rapidement.

2 Dataset

2.1 Présentation des données

Afin de réaliser ce travail, nous avons besoin d'une base de donnée adéquate, nous présentons celle-ci dans cette sous-section. La base de données est constituée d'informations sur les clients d'une compagnie d'assurance nommée Wasa, entre 1994 et 1998. Les véhicules assurées sont composés uniquement de motos. Il est difficile d'obtenir des données plus récentes à cause des clauses de confidentialité des assurances. La version originelle de ce jeu de données est utilisée dans une étude cas du livre "Non-life insurance pricing with GLM", écrit par Ohlsson et Johansson, et celle-ci est disponible sur le site web du livre[3]. Le jeu de données est constitué de 64505 observations des 9 variables suivantes :

- OwnersAge : L'âge du conducteur.
- Gender : Le sexe du conducteur.
- Zone : Variable catégorielle représentant la zone dans laquelle le véhicule est conduit..
- Class : Variable catégorielle représentant la classe du véhicule. Les classes sont assignées dans une des 7 catégories selon le ratio : $EV = \frac{kW \times 100}{kg+75}$.
- VehiculeAge : L'âge du véhicule en années.

[1] Apprentissage automatique : Méthodes statistique permettant à un ordinateur d'apprendre à résoudre un problème donné à l'aide d'une base de données pertinente.
[2] La plupart de ces algorithmes sont en fait assez peu récents mais étaient difficilement applicables en raison d'une puissance de calcul trop faible, citons par exemple le perceptron, élement de base des réseaux de neurones qui a été inventé dès 1957.
[3] http://staff.math.su.se/esbj/GLMbook/case.html

- BonusClass : Le bonus du conducteur, un nouveau conducteur commence à 1 et sera incrémenté à chaque année complète passée dans la compagnie sans sinistre déclaré, jusqu'à un maximum de 7.
- Duration : Le nombre d'année passées dans la compagnie.
- NumberClaims : Le nombre de sinistres.
- ClaimCost : Le coût des sinistres.

La variable Zone est décrite dans la table 1.

Table 1: Descriptions des variables catégorielles

Variable	Classe	Description
Zone géographique	1	Parties centrales et semi-centrales des trois plus grandes villes de Norvège.
	2	Banlieues et villes moyennes.
	3	Petites villes (à l'exception de celles des catégories 5 et 7).
	4	Villages (à l'exception de ceux des catégories 5 et 7).
	5	Villes du nord de la Suède.
	6	Campagnes du nord de la Suède.
	7	Gotland (Grande île).

2.2 Analyse exploratoire

Maintenant que les différentes variables ont été brièvement présentées et leur fonction plus claire, nous allons maintenant passer à l'analyse exploratoire de celles-ci. Le but de cette analyse est de mieux comprendre les données qui serviront à entraîner les différents algorithmes, ainsi que de repérer d'éventuelles anomalies. Durant l'analyse exploratoire d'une base de données, il est important de regarder la distribution des variables, celle-ci nous donne beaucoup d'informations quant aux données.

Sur ces deux premières figures nous observons respectivement les distributions des variables Gender et OwnersAge. La première chose que nous pouvons observer est la grande disparité entre le nombre d'hommes de et femmes clients de l'assurance. Notre jeu de données est composé d'hommes pour la grande majorité. Pour ce qui est de la variable OwnersAge, nous constatons que les valeurs sont réparties entre 16 et 92 ans, avec deux "pics" vers 25 et 45 ans. Les valeurs maximales et minimales de nos variables continues seront importante pour la suite, car elles sont nécessaires afin d'appliquer une normalisation des données, qui sera discutée plus tard dans ce travail. Ci dessous les distributions des variables VehiculeAge et Zone :

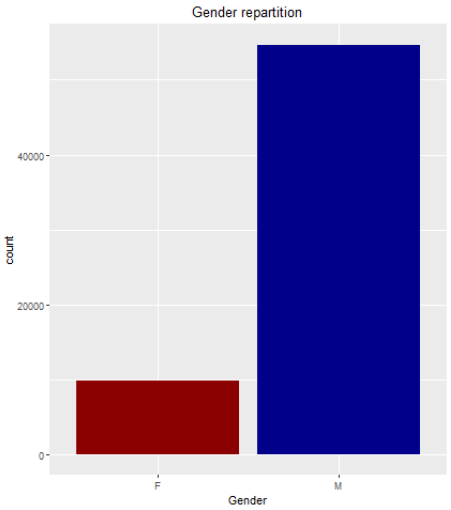

Figure 1: Distribution Gender

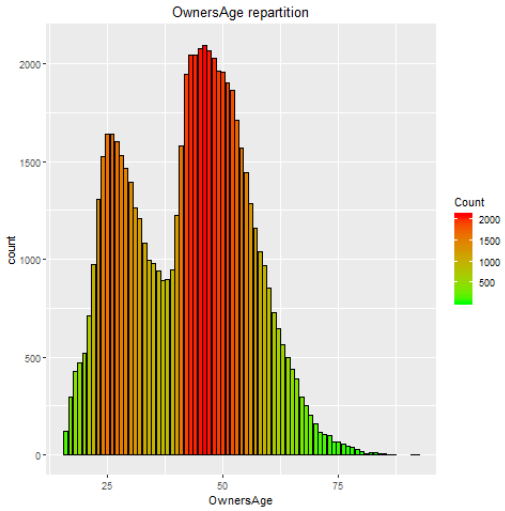

Figure 2: Distribution OwnersAge

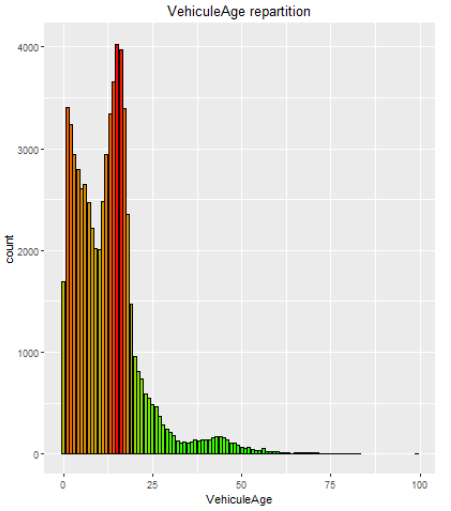

Figure 3: Distribution VehiculeAge

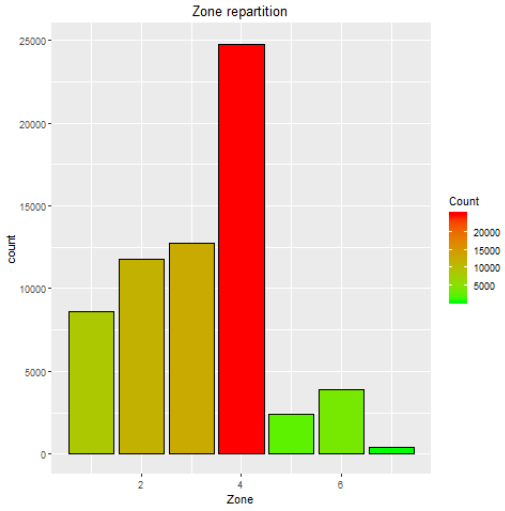

Figure 4: Distribution Zone

La distribution de vehicule est indiquée en années, et nous pouvons donc observer que si la plupart des véhicules assurés ont moins de 20 ans, un grand nombre de ceux-ci sont bien plus vieux, avec comme maximum 99 ans. Il est possible que ce véhicule soit considéré comme outlier et il sera reconsidéré dans la partie preprocessing.
La distribution de la variable Zone est intéressante, elle nous révèle que la plupart des véhicules assurés sont conduits dans des villages, mais aussi que très peu d'entre eux le sont dans le Gotland. Ceci n'est pas surprenant puisque la population de la Suède est d'environ 10 millions d'habitants, pour seulement

60.000 habitants la région du Gotland. Les classes 5 et 6 sont elles aussi minoritaires, cela était aussi à prévoir puisque ces catégories représentent le nord de la Suède alors que la plupart de la population vis dans le sud du pays.

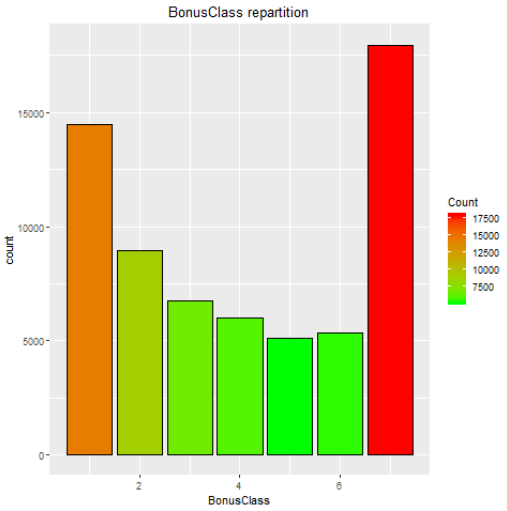

Figure 5: Distribution BonusClass

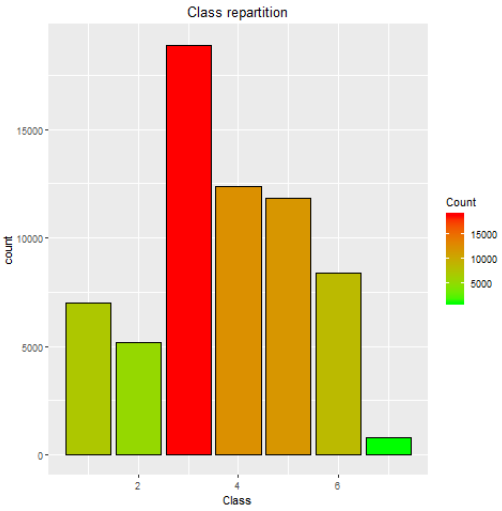

Figure 6: Distribution Class

Les classes de bonus les plus fréquentes sont les classes 1 et 7, respectivement le minimum (entrée dans la compagnie d'assurance) et le maximum (client fidèle depuis sept années au minimum).
Pour la variable Class nous observons qu'assez peu de véhicules appartiennent à la catégorie la plus puissante. En fait, la plupart des véhicules se situent dans les classes 3, 4 et 5, ce qui montre que les vehicules les moins puissants et les plus puissants sont minoritaires. Les graphiques suivants nous montrent la répartitions des sinistres par bonus et par classe de véhicule.

Il parait logique de supposer que plus un client a un bonus élevé, moins celui-ci causera d'accident, puisque le bonus monte uniquement si le client parvient à compléter une année sans causer d'accident. Cependant en observant la figure 7, il apparait que la majorité des cas d'accidents sont déclarés par des clients appartenant aux classes 3 à 6 de bonus. Ce qui est d'autant plus étonnant lorsque l'on associe ce résultat avec la distribution de la variable Class (figure 5) : les classes 3 à 6 sont celles contenant le moins d'utilisateurs. Les clients appartenant à la classe de bonus 7 semblent cependant causer très peu d'accidents malgré le fait qu'ils soient la classe de bonus majoritaire.
Les résultats de la figure 8 sont moins surprenant : plus un véhicule est puissant, plus le risque d'accident sera important. Les déviations de cette régle par les classes 3 et 7 sont expliquées par la distribution de la population à travers les différentes classes (figure 6), ainsi il y a peu d'accidents pour les véhicules de classe 7 simplement car ceux-ci sont peu nombreux, une conclusion similaire peut être énoncée pour la classe 3.

La répartition des sinistres par sexe indique que les hommes sont plus sucep-

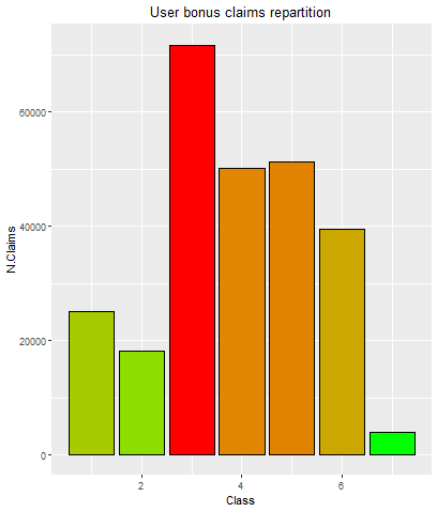

Figure 7: Bonus Claims

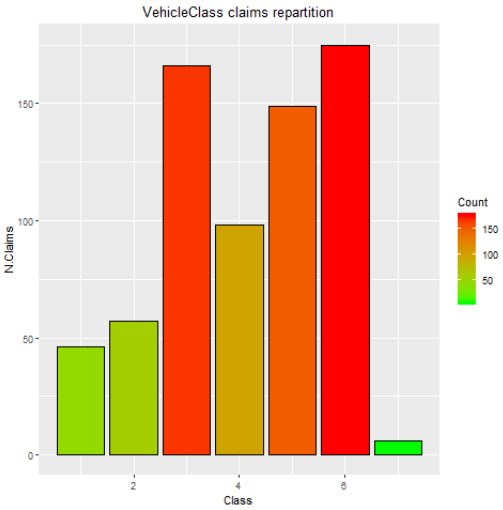

Figure 8: Class Claims

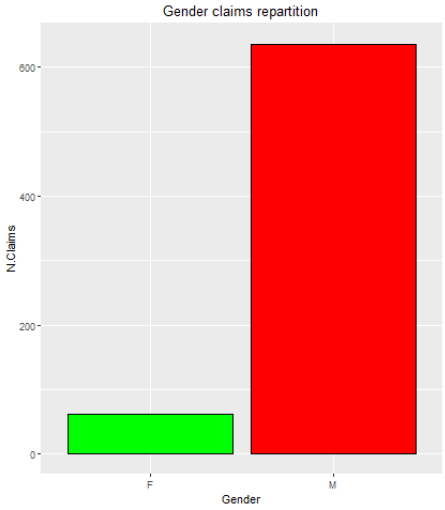

Figure 9: Gender claims

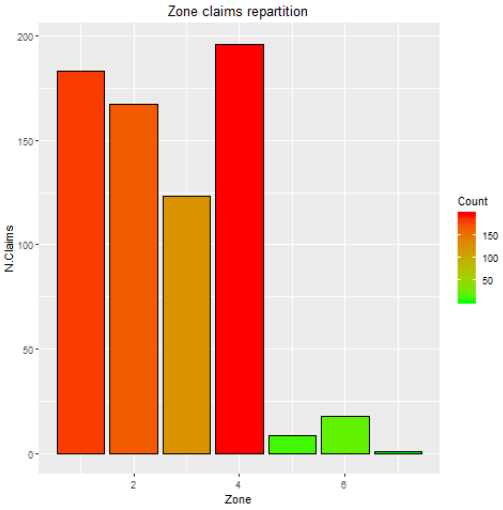

Figure 10: Zone claims

tibles de causer des accidents que les femmes. Il faut prendre en compte que les hommes sont bien plus nombreux que les femmes dans nos données, mais la conclusion ne change pas puisque la proportion de femmes est de 18%, alors que celles-ci ne causent que 8.7% des sinistres. Sur la figure 10, nous observons que les classes 1 et 3 sont celles qui déclarent le plus de sinistres. La première classe étant plutôt minoritaire (figure 4) nous observons que les personnes habitant dans les parties centrales et semi-centrales de Norvège semblent causer bien plus d'accidents que les autres catégories, la même conclusion peut être faite pour la deuxième classe (Banlieues et villes moyennes). Pour ce qui est de nord de la Suède, nous constatons l'inverse puisque ceux-ci semblent causer assez peu d'accidents.

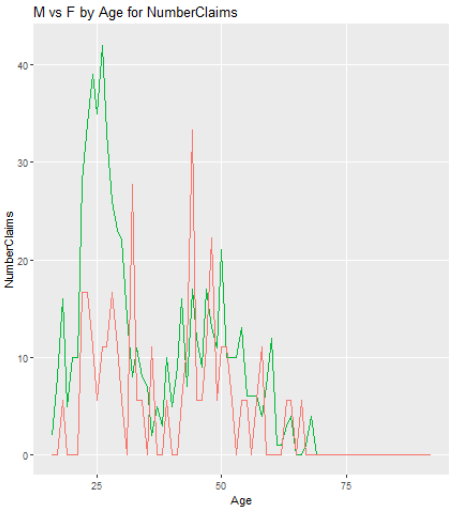

Figure 11: Gender claims by Age

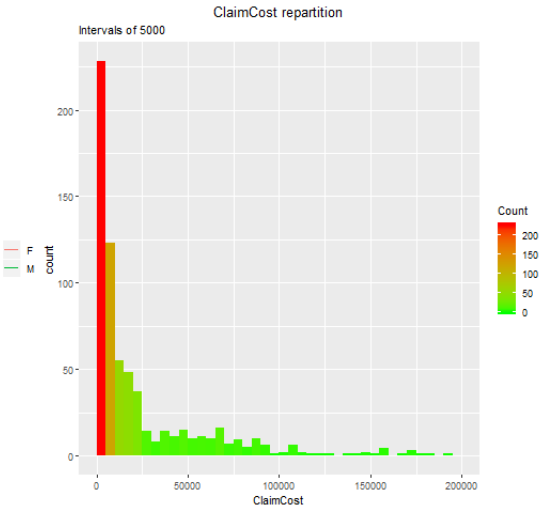

Figure 12: Distribution ClaimCost

La figure 11 a pour but de mieux comprendre cette disparité entre les hommes et les femmes dans la déclaration des sinistres. Une comparaison par âge permet de mettre en évidence une information importante. En effet, si les hommes semblent bien causer plus d'accidents que les femmes jusqu'à la treintaine, cette tendance s'égalise passé ce cap. Enfin, la figure 12 nous informe de la distribution de la variable ClaimCost, représentant le coût total des sinistres d'un client. Il est important de noter qu'il s'agit bien de la somme de tous les sinistres déclarés du client, ainsi si une personne a reçu une compensation de l'assurance pour plusieurs sinistres, la variable contient la somme du coût de tous ces sinistres déclarés. Il semble que la grande majorité des clients ne dépassent pas 5000€ de compensation venant de l'assurance (Le graphique ne montre que les valeurs non-nulles, ainsi les clients n'ayant jamais déclaté de sinistres ne sont pas comptés dans l'intervalle $[0, 5000]$. La distribution de cette variable montre bien qu'assez peu de clients dépassent les 20.000€ de compensation, mais certains peuvent monter à près de 200.000€.

2.3 Preprocessing

Le preprocessing des données est une étape très importante lorsque nous travaillons avec ce type de données et celui-ci n'est pas à négliger. Il consiste a préparer les données pour les algorithmes, afin que ceux-ci puissent fonctionner de manière optimale. Dans ce travail, plusieurs méthodes de preprocessing ont été utilisées.

2.3.1 Variables binaires (Dummy variables)

Certaines de nos variables sont catégorielles, et plus particulièrement sont des variables non-ordinales. Cela signifie que les différentes catégories ne peuvent pas être ordonnées, il s'agit purement de l'assignation d'un client à un groupe

donné. Ce type de variable ne peut pas être encodé tel quel et doit être transformé en une série de variables binaires. La conversion d'une variable catégorielle en variable binaire consiste à créer $n-1$ nouvelles variables dites "binaires" et qui prendront pour valeur 1 si le client appartient à cette catégorie, 0 sinon. Par exemple si x_{ij} représente la variable associée au client i et à la variable j, nous pouvons écrire :

$$x_{ij} = \begin{cases} 0 \text{ if } x_i \notin j \\ 1 \text{ if } x_i \in j \end{cases}$$

Les variables Gender, Zone, Class, et BonusClass peuvent être transformées de la sorte. Ce travail sera effectué avec le package caret[4] de R. La fonction incluse dans ce package crée n variables binaires, la dernière sera donc supprimée car celle-ci ne représente aucune information. Prenons pour exemple la variable Gender. Si $x_{iGender} = 1$ alors le client est une femme. Cependant nous ne devons pas créer une deuxième variable binaire puisque si $x_{iGender} = 0$ nous pouvons en déduire que le client est un homme. Ceci explique pourquoi nous n'avons besoin que de $n-1$ variables binaires pour représenter toute l'information de n catégories appartenant à une variable catégorielle.

2.3.2 Normalisation

Les données quantitatives ne doivent pas être transformées en variables binaires (bien que celles-ci peuvent être converties en intervalles puis en variables binaires, ce ne sera pas le cas dans le cadre de ce travail). Cependant cela ne signifie pas qu'aucune méthode de preprocessing ne doit être appliquée à ces données. La normalisation des données quantitatives aide certains algoithmes à converger plus rapidement et peut même parfois améliorer leur précision. Cette normalisation des données est particulièrement importante pour les réseaux de neurones, car l'optimisation de ceux-ci est basée sur les résultats des fonctions d'activation[5] utilisées. Celles-ci varient rapidement entre 0 et 1 (ou -1 et 1) et sans la normalisation des données leurs réponses aux valeurs extrêmes seraient localisées dans les extrémités, ce qui peut empêcher l'algorithme de converger dans les cas les plus graves.
Plusieurs méthodes peuvent être utilisées afin de procéder à la normalisation des données. Cependant nous ne décrirons que celle qui sera utilisée dans le cadre de ce travail : la normalisation Min Max.
Elle consiste (pour une variable j) à convertir l'ensemble des données sur l'intervalle [0,1]en se servant de $Max(x_j)$[6] et $Min(x_j)$[7] par la formule :

$$z_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)}$$

Cette transformation sera appliquée aux variables OwnersAge et VehiculeAge.

[4]Classification And REgression Tools : Package de machine learning pour R.
[5]Voir chapitre 3
[6]Valeur maximale prise par la variable j.
[7]Valeur minimale prise par la variable j.

2.3.3 Outliers

3 Algorithmes