REPAS : Responsabilité estimée par apprentissage statistique - Rapport final

Cédric Garcia, Vivian Viallon, Liacine Bouaoun, Jean-Louis Martin

Ifsttar Bron
25 Avenue François Mitterrand
69500 Bron

Université Lyon 1
43 Boulevard du 11 Novembre 1918
69100 Villeurbanne

REPAS : Responsabilité estimée par apprentissage statistique

Rapport final

Référence de la convention	2201039839
Contrat IFSTTAR	RP1-J16154
Date de notification	18 novembre 2016
Référent DSR	Arnaud Guenivet
Responsable du projet	Vivian Viallon
Date de livraison du rapport	15 janvier 2018
Auteurs	Cédric Garcia Vivian Viallon Liacine Bouaoun Jean-Louis Martin

Table des matières

1 Introduction

La survenue d'un accident corporel peut être qualifiée d'événement rare pour un conducteur [1, 2]. Dès lors le design d'étude privilégié en épidémiologie pour estimer des facteurs de risque est l'étude cas-témoins, les cas étant les accidentés et les témoins les conducteurs qui n'ont pas eu d'accident sur une période donnée. Pour des facteurs de risque transitoires (comme la conduite sous influence ou l'usage du téléphone au volant), cela implique une comparaison des usagers accidentés (représentant les cas), avec d'autres usagers présents sur la route dans la période de temps où l'accident est survenu (représentant les témoins). Pour l'étude de facteurs de risques stables (par exemple l'âge, le sexe ou la santé de l'usager), la comparaison doit être effectuée entre des usagers accidentés et des conducteurs ayant le même niveau d'exposition [3]. Il est cependant difficile de mesurer ces facteurs de risques dans un contrôle routier lors de telles études, comme par exemple l'usage du téléphone au volant ou la distraction du conducteur. De plus certains conducteurs sollicités peuvent refuser de coopérer de façon différente selon qu'ils s'estiment être en faute ou pas, comme pour un contrôle de conduite sous influence (alcool ou stupéfiants). Il est en fait quasiment impossible de connaître le niveau d'exposition à un risque d'accident dans une population de conducteurs.

Une approche a été développée pour résoudre les problèmes liés aux études cas-témoins : l'analyse en responsabilité des usagers accidentés. Cette analyse repose sur la comparaison des usagers responsables avec les usagers non responsables. Cette technique permet d'estimer le risque d'être responsable d'un accident pour chaque usager impliqué à partir des seules bases de données d'accidents de la route. Cette approche fait l'hypothèse que l'ensemble des usagers accidentés non responsables constitue un échantillon représentatif de l'ensemble des conducteurs [4, 5]. Cette hypothèse se base sur le fait que les usagers responsables d'accidents ne choisissent pas intentionnellement quels conducteurs vont être impliqués, et qu'en conséquence, on peut supposer que tous les conducteurs non responsables ont la même probabilité d'être impliqués dans un accident [6].

La validité de cette méthode suppose d'être en mesure d'estimer très précisément la responsabilité de chaque acteur impliqué dans un accident. Or ceci nécessite de disposer d'une description précise des circonstances de l'accident. Il est bien entendu que cette responsabilité n'est pas définie dans un sens juridique : un usager est considéré responsable s'il contribue, voire déclenche l'accident, typiquement par une manœuvre inappropriée (circulation en sens interdit, passage au feu rouge, etc.) ou un manquement (freinage tardif, oubli d'allumage des feux de croisement la nuit ou dans un tunnel, etc.). Il est par conséquent essentiel que la définition de la responsabilité repose directement sur ces comportements et non sur leurs causes que sont, par exemple, l'inexpérience du conducteur, l'alcool, l'usage du téléphone au volant, etc. Dans le cas contraire, les effets de ces facteurs sur le risque d'être responsable d'un accident de la route se trouveraient largement surestimés.

En France, les forces de l'ordre rédigent un procès-verbal pour tout accident corporel de la circulation routière pour lequel ils ont été sollicités. Une partie de ces procédures est informatisée en routine dans une base de données (dénommée BAAC). Elle contient diverses informations sur les accidents telles que le lieu de l'accident, les véhicules impliqués, les usagers concernés et les infractions commises. La responsabilité de chaque usager est en particulier renseignée par les forces de l'ordre. Cette variable de responsabilité de l'usager est renseignée au moment de la rédaction de la fiche, telle que l'apprécient les forces de l'ordre. Cette détermination de la responsabilité par les forces de l'ordre peut être utilisée pour conduire les analyses en responsabilité. Cependant, les critères utilisés dans l'attribution de la responsabilité ne sont pas

clairement établis et la validité de cette responsabilité n'est donc pas garantie. En effet on peut craindre que les forces de l'ordre aient tendance à adopter un point de vue trop légaliste pour déterminer la responsabilité des usagers, ce qui n'est pas cohérent avec la définition de la responsabilité évoquée précédemment. Dans des conditions idéales, une détermination fiable de la responsabilité des usagers requiert un expert entraîné pour analyser la responsabilité des usagers à partir d'une description très détaillée de l'accident. Un tel travail a été accompli dans le cadre du projet VOIESUR (Véhicule Occupant Infrastructure Etudes de la Sécurité des Usagers de la Route) (http ://www.agence-nationale-recherche.fr/ ?Projet=ANR- 11-VPTT-0007). Ceci a permis de disposer d'un critère de responsabilité fiable (au sens « contributeur ») et le plus objectif possible (basé sur les faits). La responsabilité déterminée dans la base VOIESUR est ainsi celle que l'on estime la plus fiable pour effectuer les analyses en responsabilité.

L'objectif principal de ce travail est de prédire la responsabilité établie par les experts (considérée ici comme le gold standard) à partir des informations explicites enregistrées en routine par les forces de l'ordre, pour différentes configurations d'accidents corporels. Pour cela nous évaluerons les performances de différentes méthodes d'apprentissage statistique. Pour reproduire au mieux la détermination des experts, la prédiction de la responsabilité « expert » sera effectuée à base de modèles d'apprentissage statistique. Plusieurs méthodes d'apprentissage seront mises en concurrence, et des méthodes de cross validation seront utilisées pour se garantir d'un éventuel overfitting. L'objectif final est de disposer d'une responsabilité déterminée selon des règles explicites et à travers un processus s'appuyant sur les données réelles observées.

2 Matériel et méthodes

2.1 Base VOIESUR

Dans le cadre du projet VOIESUR, une base de données a été constituée à partir de l'analyse complète et de la codification minutieuse des procédures de Police-Gendarmerie numérisées et centralisées par un organisme (TransPV) pour mise à disposition auprès des sociétés d'assurance concernées. Les services de recueil ont été directement contactés en cas de manque d'éléments manquants importants, tels que les plans de l'accident, les photos des véhicules impliqués et les bilans lésionnels. La collecte a porté sur tous les accidents mortels survenus sur le territoire métropolitain ainsi qu'un vingtième des accidents corporels pour l'année 2011. La base ainsi constituée inclut 7846 accidents.
La variable de responsabilité « expert » renseignée dans VOIESUR (notée *crespo*) peut prendre les valeurs suivantes :

1 – l'usager est totalement responsable de l'accident,

2 – l'usager est plutôt responsable de l'accident,

3 – la responsabilité de l'usager dans l'accident est partagée,

4 – l'usager est plutôt non responsable de l'accident,

5 – l'usager est totalement non responsable de l'accident.

2.2 Types d'accidents étudiés

Pour estimer la part de responsabilité d'un conducteur sur la survenue d'un accident, il est indispensable de considérer simultanément le comportement des autres usagers (conducteurs, cyclistes ou piétons) impliqués. Or les informations traduisant le comportement des usagers impliqués varient selon le type d'usager. Ainsi l'information concernant l'excès de vitesse n'est pas pertinente pour un cycliste, comme le changement de file n'a pas de sens pour un piéton.

La responsabilité va ainsi être déterminée pour les trois types d'accidents les plus fréquents

— Configuration 1 : accidents impliquant deux véhicules motorisés ou plus de deux,

— Configuration 2 : accidents impliquant un véhicule motorisé et soit un cycliste, soit un piéton,

— Configuration 3 : accidents n'impliquant qu'un véhicule motorisé.

Les accidents n'impliquant que des cyclistes ou un cycliste et un piéton ne vont donc pas être considérés.

2.3 Variables

En prenant le codage de la responsabilité « expert » définie dans 2.1, on définit le groupe des responsables comme celui des conducteurs pour lesquels la responsabilité est égale à 1 ou 2. Ce choix d'inclure les "plutôt responsables" provient du raisonnement suivant : alors qu'un accident se produit souvent à cause de la conjonction de plusieurs facteurs, la suppression de l'un d'entre eux suffit, la plupart du temps, pour que l'accident ne se produise pas. En d'autres termes, on considère que l'accident n'aurait pas eu lieu si l'usager n'avait pas fait ce qui a poussé l'expert à le déclarer plutôt responsable. Le groupe des non responsables est constitué des conducteurs pour lesquels la responsabilité est égale à 4 ou 5. Les conducteurs pour lesquels la responsabilité est égale à 3 sont les moins fréquents dans l'échantillon et peuvent poser problème dans la prédiction de la responsabilité par les modèles statistiques, c'est pourquoi il

a été décidé de les retirer de l'échantillon. On note Y la version binaire de la responsabilité « expert » définie comme :

$$Y = \begin{cases} 1 \text{ si } crespo \in \{1,2\}, \\ 0 \text{ si } crespo \in \{4,5\}. \end{cases}$$

C'est cette nouvelle variable que l'on va essayer de prédire.

Il faut également identifier l'ensemble des variables qui peuvent prédire la responsabilité. Dans la section 1 il a été vu que des actions inappropriées peuvent rendre le conducteur responsable. La base VOIESUR comporte les variables renseignées par les forces de l'ordre, notamment les infractions. Quelques-unes d'entre elles sont sélectionnées car elles indiquent des actions inappropriées qui peuvent entraîner l'accident, en particulier : conduite en sens interdit, vitesse excessive, refus de priorité... Ainsi que déjà évoqué précédemment, nous n'incluons pas des possibles causes de ces comportements, comme l'alcoolémie, l'âge du conducteur ou le nombre d'années d'expérience... Les autres variables explicatives retenues sont celles qui peuvent avoir une influence sur la responsabilité. Les variables explicatives choisies sont soit numériques, soit binaires. Parmi elles, soit $(W_1, \ldots, W_p)$ l'ensemble des variables décrivant les caractéristiques de l'accident (conditions météorologiques, état de la surface de la route...) ; les valeurs de celles-ci doivent être les mêmes pour tous les usagers impliqués dans un même accident, et $(Z_1, \ldots, Z_q)$ les variables qui dépendent de l'usager (dernière manœuvre effectuée, position du point d'impact du véhicule...). A noter que l'ensemble de variables explicatives est différent selon la configuration d'accident précédemment définie.

2.3.1 Accident impliquant un seul usager motorisé

Pour ce type d'accident, chaque observation dans la base de données correspond exactement à un accident, le comportement de l'usager n'est pas influencé par les autres usagers. Il n'y a pas de changement effectué sur les variables explicatives.

2.3.2 Accidents impliquant au moins deux usagers motorisés

Lorsqu'au moins deux usagers motorisés sont impliqués, le comportement de l'un d'entre eux est influencé par celui des autres. Cette interaction doit apparaître dans les variables explicatives, afin de déterminer la responsabilité d'un usager. Pour ce faire, on change les variables $(Z_1, \ldots, Z_q)$ en $(S_1, \ldots, S_q)$, de telle manière à ce que pour un accident impliquant les usagers $A_1, \ldots, A_r$ on définit :

$$S_l(A_i) = Z_l(A_i) - \max_{j \neq i} Z_l(A_j), \quad 1 \leqslant i \leqslant r, 1 \leqslant l \leqslant q.$$

En particulier pour un accident impliquant un usager A et un usager B, $S_l(A) = Z_l(A) - Z_l(B)$, pour l entre 1 et q. Pour les accidents impliquant au moins deux véhicules motorisés, il peut y avoir des hétérogénéités de l'influence des variables explicatives, selon le nombre exact de véhicules impliqués. On pense notamment qu'il peut y avoir une différence entre deux véhicules impliqués, et trois véhicules ou plus. Pour mesurer cette différence, on change le codage des variables explicatives. Pour une observation quelconque, soit $\widetilde{x}^T = (W_1, \ldots, W_p, S_1, \ldots, S_q)$ le vecteur des variables explicatives, on définit :

$x^T = (\widetilde{x}^T, 0^T)$ si 2 véhicules sont impliqués,
$x^T = (\widetilde{x}^T, \widetilde{x}^T)$ si 3 véhicules ou plus sont impliqués.

Avec ce codage des variables explicatives, le modèle logit sera le suivant :

$$\text{logit}[P(Y = 1|X = x)] = \alpha + \widetilde{x}^T\beta + \mathbb{1}[\text{Au moins 3 véhicules}] \times \widetilde{x}^T\gamma.$$

On effectue le même codage pour les accidents impliquant un cycliste ou un piéton, et un véhicule motorisé :

$x^T = (\widetilde{x}^T, 0^T)$ si un piéton est impliqué,
$x^T = (\widetilde{x}^T, \widetilde{x}^T)$ si un cycliste est impliqué.

2.4 Modèles statistiques

On a défini Y comme la variable binaire de la responsabilité « expert ». On cherche à prédire les valeurs de Y avec notre ensemble de variables explicatives X. C'est un problème de classification. On va mettre en concurrence trois méthodes différentes pour tenter de prédire Y :

- Régression logistique avec sélection des variables par lasso
- Forêts aléatoires
- Boosting

2.4.1 Régression logistique avec pénalisation lasso

La régression logistique est une méthode de classification d'une variable binaire, permettant de construire un score. Si Y est une variable binaire à expliquer, et X le vecteur des variables explicatives, alors :

$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \alpha + x^T\beta, \text{ ce qui donne :}$$

$$P(Y=1|X=x) = \frac{\exp(\alpha+x^T\beta)}{1+\exp(\alpha+x^T\beta)} = \frac{1+\tanh\left(\frac{\alpha+x^T\beta}{2}\right)}{2}.$$

Cependant le nombre de variables explicatives est élevé. Le modèle de régression logistique risque de retenir trop de variables, et ainsi le modèle risque d'effectuer du surapprentissage. C'est pourquoi une sélection initiale de variable est effectuée, à l'aide d'une pénalisation LASSO (Least Absolute Shrinkage and Selection Operator). Si $L(\beta)$ est la vraisemblance du modèle, la pénalisation LASSO revient à maximiser $L(\beta) - \lambda||\beta||_1$ [7]. Pour une séquence de λ générée par la méthode, le modèle LASSO sélectionne ainsi les variables. Une régression logistique est alors effectuée en ne gardant que ces variables sélectionnées. De cet ensemble de modèles de régression logistique, le choix est effectué par le critère AIC (Akaike Information Criterion). Il se calcule par $AIC = -2L(\beta) + 2k$ où k est le nombre de paramètres du modèle. Le modèle choisi est celui qui présente le AIC le plus faible. Le choix de ce critère pour sélectionner le modèle est justifié par le fait qu'il permet d'avoir le modèle donnant les prédictions les plus précises.

2.4.2 Arbres de décision

Les arbres de décision sont également des méthodes classiques de classification. Le choix de ces méthodes pour la prédiction de la responsabilité vient du fait qu'on estime que le principe d'analyse la responsabilité d'un usager dans un accident est similaire à celui de l'établissement d'un arbre de décision ; si le conducteur a fait une faute de conduite, telle que passer un feu rouge, il n'est pas nécessaire de voir plus loin, le conducteur est certainement à l'origine de l'accident. Si en revanche un conducteur n'a pas commis de faute, il faut voir ses autres actions

en détail mais dans la plupart des cas, ce conducteur n'a rien fait pour que l'accident survienne. La première méthode d'arbre de décision sélectionnée est le *random forest*, cet algorithme d'arbre de décision présente l'avantage de ne pas effectuer de surapprentissage, contrairement à la plupart des méthodes d'arbres de décision. En effet l'algorithme du *random forest* se construit sur des prédictions out-of-bag [8]. La deuxième méthode d'arbre de décision choisie est le *boosting*. Son principe est de partir d'un modèle faible pour le transformer en un modèle précis. Comparé à l'algorithme de *random forest*, le *boosting* donne de meilleures prédictions sur l'ensemble d'apprentissage, mais présente en retour un risque de surapprentissage.

3 Résultats obtenus

3.1 Accidents impliquant au moins deux véhicules

On veut mesurer l'efficacité des prédictions obtenues par les modèles lorsqu'on applique sur de nouvelles données. Pour procéder ainsi on met en place une méthode de validation croisée pour la régression logistique et le *boosting*. L'échantillon de 7597 observations est divisé en 5 sous-ensembles. Les modèles sont construits sur l'union de 4 sous-ensembles, et on teste leurs prédictions sur le sous-ensemble restant. Étant donné qu'il y a 5 sous-ensembles, cela permet d'obtenir 5 estimateurs, un pour chaque ensemble test. L'estimateur final est obtenu en calculant la moyenne des 5 précédents. Le random forest ne nécessite pas de validation croisée, car il l'effectue par lui-même en utilisant les prédictions out-of-bag. On choisit de mesurer 5 critères, si on note $\widehat{Y}$ la prédiction d'un modèle :

- La précision, définie comme la proportion d'observations où la responsabilité observée correspond à celle prédite par le modèle
- La sensibilité, définie par $Sn = \mathbb{P}(\widehat{Y} = 1 | Y = 1)$
- La spécificité, définie par $Sp = \mathbb{P}(\widehat{Y} = 0 | Y = 0)$
- L'aire sous la courbe ROC (AUC)
- Le coefficient kappa de Cohen

Les résultats obtenus sont présentés dans la table 1.

	Régression logistique	Forêts aléatoires	Boosting
Précision	0,869 [0,862 ; 0,877]	0,864 [0,856 ; 0,871]	0,869 [0,862 ; 0,876]
Sensibilité	0,887 [0,877 ; 0,898]	0,812 [0,799 ; 0,825]	0,837 [0,825 ; 0,849]
Spécificité	0,853 [0,843 ; 0,864]	0,909 [0,900 ; 0,918]	0,898 [0,889 ; 0,906]
AUC	0,936 [0,931 ; 0,942]	0,932 [0,927 ; 0,938]	0,936 [0,931 ; 0,940]
Kappa de Cohen	0,739 [0,724 ; 0,754]	0,725 [0,710 ; 0,741]	0,737 [0,722 ; 0,751]

TABLE 1 – Résultats obtenus par validation croisée pour la régression logistique et le boosting, et par prédictions out-of-bag pour les forêts aléatoires

Ces résultats indiquent une bonne performance des modèles, ils peuvent être utilisés pour prédire la responsabilité pour ce type d'accident. Comme les résultats sont très similaire pour chaque modèle, on privilégie le modèle de régression logistique ; celui-ci apporte un critère mathématique utilisable pour évaluer la responsabilité, et donne une liste explicite des variables les plus influentes avec le calcul des rapports des côtes ajustés.

3.1.1 Influence des facteurs de risque

L'utilisation principale de la responsabilité est de déterminer le risque d'être responsable selon divers facteurs de risque. Une manière de vérifier la fiabilité du modèle construit est d'estimer

le risque associé d'être responsable selon divers facteurs de risque, en comparant les résultats obtenus avec la responsabilité « expert » comme référence. On compare ainsi la pertinence de la responsabilité des forces de l'ordre et du modèle de régression logistique.

Les facteurs de risque choisis sont le taux de concentration d'alcool en grammes par litre de sang (taux inférieur à 0,5 comme référence), l'âge (19 ans ou moins comme référence), le sexe (homme en référence), la catégorie socioprofessionnelle (les cadres et ouvriers sont choisis comme référence), l'état du permis de conduire (permis valide choisi comme référence) et le type d'usager (voiture en référence). Tous ces facteurs de risque choisis ont été codés de manière qualitative.

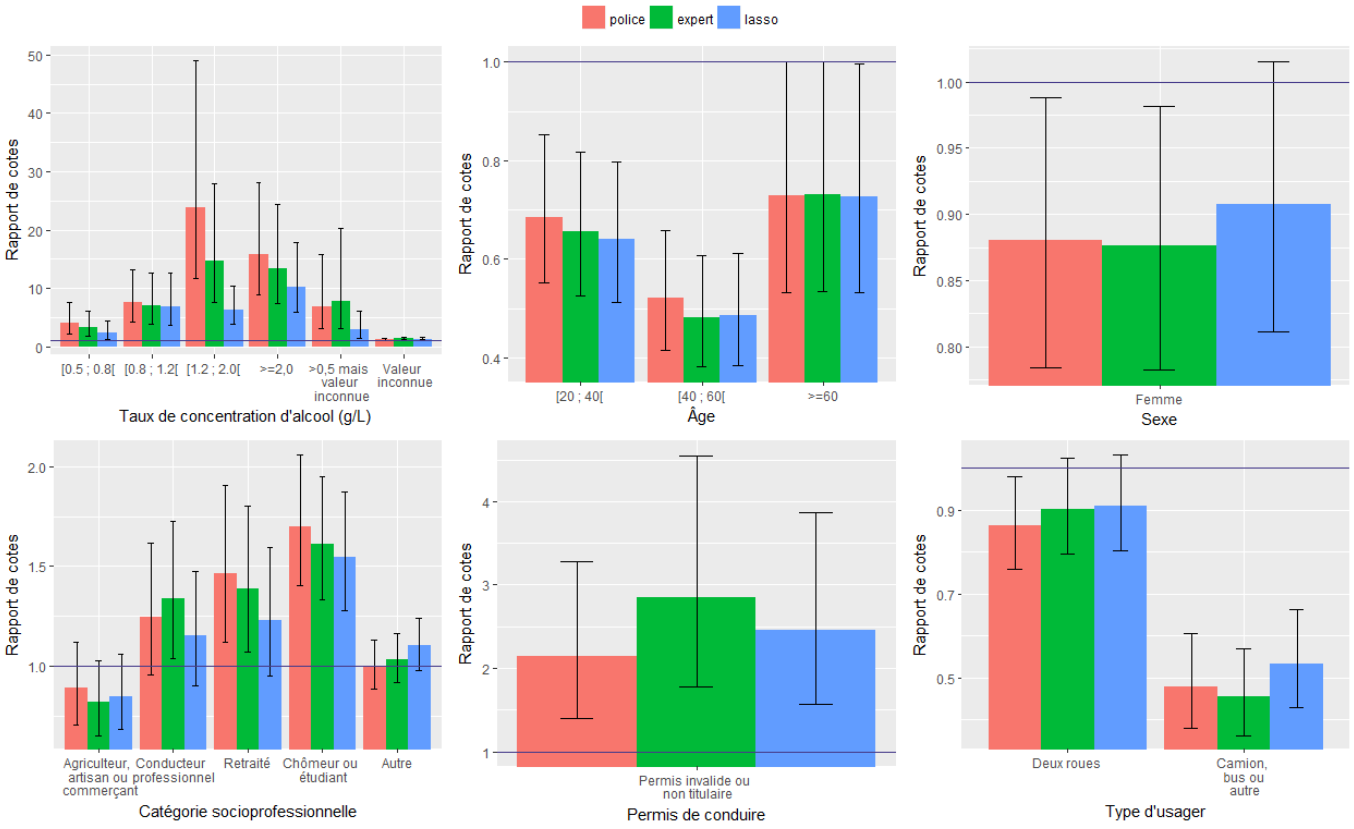

FIGURE 1 – Rapports de cotes ajustés des différentes responsabilités en fonction de divers facteurs de risque

On remarque que la police surestime les effets de l'alcool sur la responsabilité lorsque le taux d'alcool est compris entre 1,2 et 2,0 g/L. On remarque que pour les autres facteurs de risque ; les rapport de cotes sont très similaires pour les 3 indicateurs. Cependant, le sexe n'a significativement pas d'influence sur la responsabilité pour le modèle de régression logistique, et les usagers en véhicule à deux roues ont significativement moins de chance d'être responsable selon la police. Ces résultats semblent indiquer que la plupart des facteurs de risque ont une influence similaire sur la responsabilité des forces de l'ordre comparée à celle des experts.

3.1.2 Construction du score

Le modèle de régression logistique permet l'élaboration d'un score d'évaluation de la responsabilité et indiquant l'influence des variables explicatives. Le score noté R se calcule par la formule suivante :

$$R = \alpha + x^T \beta.$$

Ce score permet de calculer la probabilité p pour un usager d'être responsable d'un accident, par la formule :

$$p = \frac{\exp(R)}{1+\exp(R)} = \frac{1+\tanh\left(\frac{R}{2}\right)}{2}.$$

On fixe le seuil de probabilité à 0.5 pour différencier les responsables des non responsables. Ainsi si $R > 0$ l'usager est responsable, sinon il est non responsable.

	Variable	Valeur du coefficient β pour un accident impliquant deux véhicules	Valeur du coefficient β pour un accident impliquant au moins trois véhicules
		$\alpha = 0,0231$	$\alpha = 0,185$
Conditions extérieures de l'accident	Route à sens unique	0	$-0,688$
	Route à chaussées séparées à 2x2 voies ou route bidirectionnelle à quatre voies	0	$-0,553$
	Accident ayant eu lieu en agglomération	0	$-0,466$
	Accident ayant eu lieu dans un pont, tunnel ou souterrain	0	$-0,466$
	Route nationale	0	$-0,361$
	Forte pluie	0	$0,629$
	Brouillard, neige ou tempête	0	$0,979$
Actions du conducteur ou configuration de l'accident	Véhicule heurté à l'arrière	$-1,35$	$-0,730$
	Véhicule immobile	$-0,511$	
	Vitesse excessive	0	$0,245$
	Manœuvre d'évitement	0	$0,576$
	Véhicule heurté à l'avant	0	$1,19$
	Véhicule heurté à droite	$0,0618$	
	Véhicule tournant à gauche	$0,143$	
	Obstacle mobile heurté	$0,198$	$0,430$
	Dépassement d'un véhicule par la gauche	$0,215$	
	Obstacle fixe heurté	$0,524$	
	Véhicule déporté à gauche	$1,10$	
	Véhicule tournant à gauche ou à droite	$1,13$	
	Insertion du véhicule	$1,47$	
	Dépassement d'un véhicule	$1,53$	
	Entre deux files, demi-tour ou marche arrière	$1,64$	
	Nombre de fautes commises	$1,90$	$2,22$
	Changement de voie	$2,03$	
	Véhicule déporté	$2,06$	

TABLE 2 – Coefficients attribués aux variables explicatives par le modèle de régression logistique

Il y a des variables pour lesquelles les coefficients sont égaux, peu importe le nombre de véhicules motorisées impliqués, par exemple pour la variable "Véhicule immobile". Mais d'autres variables ont un coefficient différent selon le nombre de véhicules impliqués. Les conditions extérieures de l'accident n'ont pas d'influence sur la responsabilité quand deux véhicules sont impliqués. En particulier, les mauvaises conditions météo et les mauvais états de la route n'ont pas d'influence sur la responsabilité dans un accident à deux véhicules. En revanche quand au moins trois véhicules sont impliqués, quelques-unes des caractéristiques de l'accident telles que les conditions météo et le nombre de voies ont une influence sur la responsabilité. Il est également intéressant de voir qu'être heurté à l'avant n'a pas d'influence pour deux véhicules, mais comme on pouvait s'y attendre, cette variable a une influence importante quand au moins trois véhicules sont impliqués, ce qui peut être compréhensible pour des accidents ayant eu lieu suite à de forts ralentissements associés à des trafics élevés.

La variable "nombre de fautes" est numérique, toutes les autres variables sont qualitatives, leurs valeurs pouvant être -1, 0 ou 1. Quelques-unes des variables sont liées entre elles : par exemple si un conducteur effectue un dépassement par la gauche, il effectue a fortiori un dépassement. Cela indique notamment qu'effectuer un dépassement par la gauche présente plus de risque d'être responsable que de dépasser par la droite. Les actions du conducteur ont globalement la même influence sur la responsabilité quel que soit le nombre de véhicules impliqués.

3.2 Accidents à un véhicule

Dans cette partie on s'intéresse aux accidents n'impliquant qu'un seul véhicule et rien d'autre, c'est-à-dire les accidents où un usager en voiture heurte une autre voiture sans occupant ou tout autre obstacle non humain. Dans la base de données VOIESUR, l'échantillon est constitué de 1961 observations. On peut remarquer que pour ce type d'accident, l'usager est en très grande partie responsable. En effet sur notre échantillon, les experts jugent responsable 1923 usagers, soit 98 % des usagers. Être non responsable dans un accident à un seul véhicule est ainsi un évènement rare, cela risque de poser problème pour les 3 modèles pour identifier les facteurs qui pourraient expliquer quand un usager est non responsable. En particulier il peut y avoir un effet de surapprentissage pour les modèles.

3.2.1 Prédiction de la responsabilité

Tout comme pour les accidents à deux véhicules, on essaye d'appliquer les 3 différentes méthodes pour prédire la responsabilité. On applique à nouveau une méthode de validation croisée pour estimer les différents critèes de précision. Mais compte tenu du très faible nombre d'observations correspondant aux usagers non responsables, on décide d'effectuer la validation croisée avec 2 sous-échantillons au lieu de 5. On construit les modèles de régression logistique et de *boosting* par cette méthode. Le modèle des forêts aléatoires est appliqué sur l'échantillon entier et les prédictions out-of-bag remplacent la nécessité d'une méthode de validation. On donne les résultats dans la table 3.

	Régression logistique Lasso	Forêts aléatoires	Boosting
Taux de bonnes prédictions	0,981 [0,975 ; 0,987]	0,983 [0,976 ; 0,988]	0,980 [0,976 ; 0,983]
Sensibilité	0,999 [0,996 ; 1]	0,997 [0,994 ; 0,999]	0,998 [0,998 ; 0,998]
Spécificité	0,107 [0,043 ; 0,244]	0,263 [0,088 ; 0,570]	0,036 [0,008 ; 0,149]
AUC	0,562 [0,488 ; 0,636]	0,629 [0,517 ; 0,740]	0,625 [0,625 ; 0,625]
Coefficient kappa de Cohen	0,155 [0 ; 0,325]	0,370 [0,204 ; 0,537]	0,059 [0 ; 0,159]

TABLE 3 – Indicateurs obtenus par validation croisée pour la régression logistique et le boosting, les forêts aléatoires sont basées sur les prédictions *out-of-bag*

Les résultats montrent que le modèle des forêts aléatoires est meilleur que les deux autres, mais malgré tout aucun des modèles n'est capable de bien prédire quand un usager est non responsable. Cela peut indiquer que les causes qui rendent un usager non responsable ne sont pas présentes dans la base de données BAAC. Une autre explication de ces résultats pourrait venir du fait que la responsabilité « expert » est mal codée pour ce type d'accident. En effet, si on regarde l'ensemble des usagers non responsables selon les experts, la responsabilité établie par les forces de l'ordre précise que 45 % sont responsables. Ce désaccord entre police et experts n'était pas aussi important que cela pour les accidents impliquant au moins deux véhicules motorisés.

3.3 Accidents entre un véhicule motorisé et un cycliste ou un piéton

Pour ce type d'accident, on restreint volontairement les observations aux accidents impliquant au moins un véhicule motorisé. Ce choix a été fait car les variables explicatives des BAAC sont plus adaptées à la description du comportement des conducteurs de véhicules motorisés, moins à celui d'un cycliste et encore moins à celui d'un piéton. La base de données VOIESUR comprend 1262 observations correspondant à des accidents comprenant un piéton ou un cycliste, avec un véhicule motorisé. 946 d'entre elles correspondent à des conducteurs ayant eu un accident avec un piéton, les 634 autres concernent les cyclistes. Les BAAC comprennent 2 variables renseignant des informations spécifiques au piéton : sa localisation et ses actions. Cependant les BAAC ne renseignent pas des informations importantes comme la traversée du passage piéton lorsque le feu piéton est rouge, ou le refus de priorité accordé à un train, qui pourraient expliquer pourquoi le conducteur du véhicule n'est pas responsable pour ces évènements. On implémente les 3 modèles, en ajoutant ces 2 variables pour les piétons. La même méthode de validation croisée que pour le accidents à au moins deux véhicules est appliquée.

	Régression logistique Lasso	Forêts aléatoires	Boosting
Taux de bonnes prédictions	0,769 [0,746 ; 0,791]	0,763 [0,739 ; 0,786]	0,753 [0,732 ; 0,774]
Sensibilité	0,785 [0,755 ; 0,815]	0,803 [0,771 ; 0,831]	0,758 [0,729 ; 0,786]
Spécificité	0,749 [0,715 ; 0,783]	0,716 [0,678 ; 0,751]	0,745 [0,714 ; 0,777]
AUC	0,821 [0,801 ; 0,841]	0,812 [0,788 ; 0,836]	0,817 [0,801 ; 0,833]
Coefficient kappa de Cohen	0,533 [0,488 ; 0,578]	0,521 [0,474 ; 0,568]	0,502 [0,460 ; 0,544]

TABLE 4 – Résultats obtenus par validation croisée pour la régression logistique et le boosting, et par prédictions out-of-bag pour les forêts aléatoires

Le modèle de régression logistique apparaît comme le meilleur, on peut l'utiliser pour obtenir un score, de la même manière que précédemment. Néanmoins les résultats sont un moins bons que ceux obtenus pour les accidents à deux véhicules, la précision est de "seulement" 77% (comparé à 87% pour un accident entre des véhicules motorisés), et le coefficient kappa de Cohen indique un accord modéré.

3.3.1 Construction du score

On utilise de la même manière que pour les accidents impliquant au moins deux véhicules, le modèle de régression logistique pour construire un score indiquant si un usager est responsable. Avec les mêmes notations, le score R se calcule comme :

$$R = \alpha + x^T \beta.$$

Comme le score précédent, si $R > 0$, l'usager est déclaré responsable, il est déclaré non responsable sinon. La table 5 indique la valeur des différents coefficients.

	Variable	Valeur du coefficient β si l'accident implique un piéton	Valeur du coefficient β si l'accident implique un cycliste
		$\alpha = 0,511$	
Conditions extérieures de l'accident	Route bidirectionnelle à trois voies	$-1,36$	$1,31$
	Nuit sans éclairage	$-1,01$	0
	Carrefour aménagé	$-0,581$	0
	Accident ayant eu lieu dans une intersection de la forme d'un "X"	$-0,569$	
	Régime de circulation à chaussées séparées	$-0,562$	
	Route à chaussées séparées à 2x2 voies ou route bidirectionnelle à quatre voies	$-0,555$	$-1,47$
	Accident sur un terre-plein central	0	$-0,894$
	Souterrain, tunnel ou autopont	0	$-0,749$
	Route en pente	0	$-0,487$
	Accident ayant eu lieu en agglomération	0	$-0,483$
	Accident ayant eu lieu dans un intersection de la forme d'un "T" ou "Y"	0	$-0,392$
	Conditions météorologiques mauvaises	0	$1,25$
	Giratoire	0	$2,66$
	Route à sens unique	$0,373$	
	Accident ayant eu lieu ailleurs que sur la chaussée	$1,60$	
Actions du conducteur ou configuration de l'accident	Piéton sur le passage piéton	$-1,07$	0
	Vitesse excessive	$-0,597$	
	Véhicule heurté à l'arrière	0	$-1,14$
	Véhicule qui tourne à gauche ou à droite	0	$0,289$
	Obstacle mobile heurté	0	$0,532$
	Insertion du véhicule	0	$0,834$
	Changement de direction du véhicule sans avertissement préalable	0	$0,838$
	Véhicule déporté à gauche	0	$0,901$
	Dépassement d'un véhicule	0	$1,00$
	Véhicule déporté à gauche ou à droite	0	$1,04$
	Véhicule tournant à gauche	0	$1,18$
	Refus de priorité	0	$1,46$
	Changement de file	0	$2,11$
	Piéton masqué, jouant ou courant	$0,877$	0
	Nombre de fautes commises	$1,25$	$1,41$

TABLE 5 – Coefficients attribués aux variables explicatives par le modèle de régression logistique

Le modèle de régression logistique prend en compte plus de variables que précédemment pour cette configuration d'accident, comparé aux accidents impliquant au moins deux véhicules. En particulier, lorsque l'accident se produit entre un véhicule motorisé et un cycliste, plus de conditions extérieures ont une influence sur la responsabilité que lorsque l'accident implique un

piéton. Ainsi le type, la forme de la route ainsi que les conditions météorologiques apparaissent comme des facteurs significatifs. En revanche peu d'actions du conducteur ou de configurations de l'accident sont significatives quand l'usager heurté est un piéton, seules les actions de ce dernier ainsi que la vitesse et les fautes commises par le conducteur influencent la responsabilité.

4 Discussion

L'estimation de la responsabilité présentée dans ce travail repose sur les étapes suivantes :

— La responsabilité de chaque conducteur a été déterminée par des experts au vu de l'ensemble des informations contenues dans les procédures de police, y compris les plans et les photos de l'accident, pour environ 5000 accidents corporels. La consistance entre experts pour attribuer cette responsabilité a fait l'objet du développement d'une méthode particulière [9], qui n'a identifié aucune hétérogénéité entre les experts, suggérant que les différents experts ont effectivement utilisé des règles de classement similaires pour déterminer la responsabilité des conducteurs.

— Partant de cette responsabilité "expert", trois techniques d'apprentissage supervisé ont été mises en œuvre pour prédire cette valeur de référence. Après validation croisée pour la régression logistique et le boosting, et out-of-bag pour les forêts aléatoires, les trois méthodes ont montré des performances similaires en termes de taux de bonnes prédictions, de sensibilité, de spécificité ou de reliability pour les accidents de type 1 et 2. Nous avons choisi de privilégier la régression logistique, qui permet de fournir un mode de calcul simple de la responsabilité.

— Les prédicteurs que nous avons utilisés sont des codages simples des informations présentes dans les données françaises recueillies et informatisées en routine par la police en France. Même si la plupart de ces informations sont présentes dans les données similaires recueillies dans beaucoup de pays (ref IRTAD), notre proposition est bien sûr optimisée sur nos données. Nous avons cependant évité tout sur-apprentissage sur nos données en utilisant les techniques décrites dans la section méthode.

Deux travaux ont été publiés sur le même sujet, celui de Robertson and Drummer [10], et plus récemment celui de Brubacher [3]. Un score global de responsabilité a été élaboré en affectant a priori des scores (entre 1 et 5) à chaque facteur contribuant (par exemple conducteur ne respectant pas le code de la route, score=1) ou atténuant la responsabilité du conducteur (par exemple véhicule heurté, score=5). Un score supérieur à 15 indique un non responsable, un score de 13 ou moins un responsable, les scores de 14 ou 15 sont considérés comme une responsabilité indéterminée. Dans les deux cas, les auteurs n'avaient pas, à notre connaissance, de valeur de référence pour la responsabilité. La validation de leurs scores a donc été faite a posteriori en confrontant leur score à l'avis d'experts sur un échantillon de leur population d'étude. Nous avons pu appliquer les recommandations de Roberstson and Drummer à nos données. La comparaison est cependant difficile, puisqu'elle passe par une adaptation des variables publiées à nos données, avec obligatoirement une baisse des performances due aux différences d'informations disponibles dans nos données comparées aux données australiennes qui ont servi à construire le score. Il est cependant intéressant de noter que nous obtenons une prédiction de la responsabilité correcte pour les accidents impliquant deux véhicules motorisés ou plus, mais de moins bonne qualité pour les accidents impliquant un cycliste ou un piéton. Dans les deux cas, la responsabilité déterminée par les experts est mieux prédite par notre méthode que par celle déduite de Robertson et Drummer.

Le modèle a été validé partiellement sur l'objectif d'analyse en responsabilité, en estimant les rapports des cotes obtenus pour quelques facteurs de risque, en comparant avec la responsabilité

de la police et de celle des experts. On a observé des rapports des cotes semblables, excepté pour l'alcool où des différences notables apparaissaient pour les taux de concentration d'alcool élevés. Cela pourrait être expliqué par le fait qu'a ce niveau de taux d'alcool, une simple mauvaise classification de la responsabilité à une grande répercussion sur le rapport des cotes. Brubacher a également observé une différence similaire pour des taux d'alcool élevés [3]. Notons cependant que ces écarts ne remettent pas en cause le fait qu'une alcoolémie élevée est fortement associée au risque d'être responsable d'accident corporel.

5 Conclusion

L'analyse en responsabilité permet de quantifier, sous certaines hypothèses, des facteurs de risque à partir des seules données d'accident, en se passant donc de données d'exposition, ce qui explique qu'elle soit assez largement utilisée. Les résultats obtenus dépendent en grande partie de la qualité de la détermination de cette responsabilité. Il est tout aussi important que les éléments permettant cette détermination soient explicites pour que l'interprétation des facteurs de risque identifiés soit possible.

La responsabilité des forces de l'ordre est la plus accessible dans le sens où elle est disponible pour chaque accident corporel enregistré par les forces de l'ordre. Cependant elle ne semble pas être aussi fiable que la responsabilité établie par les experts de VOIESUR, base de données disponible seulement en 2011. A l'aide de techniques d'apprentissage statistiques, nous avons élaboré un modèle statistique permettant de prédire la responsabilité du point de vue des experts de VOIESUR pour les accidents des BAAC impliquant au moins deux véhicules motorisés, et les accidents impliquant un véhicule motorisé avec un piéton ou un cycliste. Le modèle utilisé pour prédire la responsabilité semble satisfaisant notamment pour les accidents impliquant au moins deux véhicules dans la mesure où le taux de bonnes prédictions est de l'ordre de 87% sous validation croisée, avec une sensibilité et une spécificité homogènes. Cependant pour les accidents à un véhicule, on n'a pas pu construire de modèle satisfaisant car ceux-ci fournissent un taux de faux responsables trop élevé. Le faible nombre de non responsables rend difficile la construction d'un bon modèle de classification dans ce cas. Ce problème peut venir en partie du fait que la responsabilité expert est mal codée pour ce type d'accident, ou que les BAAC ne permettent pas de déterminer les causes qui rendent un usager non responsable lorsque c'est le cas.

Au final on estime que les prédictions établies par le modèle peuvent être utilisées pour effectuer des analyses en responsabilité, avec plus de fiabilité que la responsabilité évaluée par les forces de l'ordre chaque année. Le modèle se base sur les variables des données BAAC, qui sont typiques de celles renseignées dans les bases de données nationales des accidents de la route, notamment dans les données de la base européenne CARE. Le modèle peut être adapté pour être utilisé avec d'autres bases de données accidentologiques si les informations servant à la construction des scores sont similaires. Dans le cas contraire, il faut disposer de valeurs de référence de la responsabilité pour appliquer la méthodologie décrite dans le présent document.

6 Annexe

6.1 Comparaison avec la responsabilité de Robertson et Drummer

Une comparaison avec la responsabilité "expert" a montré que notre modèle donne un meilleur coefficient kappa de Cohen, comparé à celui obtenu si on compare la responsabilité "expert" avec celle de Robertson et Drummer. . Pour les accidents impliquant deux véhicules motorisés ou plus, le kappa de Cohen est égal à 0.726 [0,708 ; 0,744], proche de celui obtenu avec la prédiction par modèle logistique (voir table 1), mais il est seulement de 0,159 [0,131 ; 0,187] pour les accidents impliquant un cycliste ou un piéton, à comparer avec 0,533 obtenu avec notre prédiction (voir table 4).

6.2 Détail des méthodes

Dans toute la suite on notera p le nombre de variables explicatives, n le nombre d'observations, $x_{i,j}$ la valeur de la j-ème variable pour la i-ème observation et y_i la valeur de la réponse pour la i-ème observation.
On pose

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \text{ et } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

6.2.1 Régression logistique avec sélection de variables par pénalisation lasso

Lasso

La méthode du lasso (Least Absolute Shrinkage and Selection Operator) [7] consiste à chercher

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j^2 \right)$$

$$\text{sous la contrainte } \sum_{j=1}^{p} |\beta_j| \leqslant t.$$

Cela correspond à minimiser la somme des carrés des résidus mais avec une pénalisation de la norme ℓ_1 du vecteur $(\beta_1, \ldots, \beta_p)$.

En réécrivant ce problème d'optimisation sous contrainte dans sa forme Lagrangienne, alors cela devient :

$$\underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

λ est un paramètre positif relié à t par une relation dépendante des données. Le paramètre λ que l'on choisit est celui pour lequel le modèle présente le BIC le plus faible.

On remarque que :

- Si $\lambda = 0$ (ce qui correspond à t supérieur à la norme ℓ_1 de l'estimateur des moindres carrés), cela se ramène à une régression linéaire multiple.
- Si $\lambda \to \infty$, alors $\beta_j = 0$ pour tout $1 \leqslant j \leqslant p$.

De part la nature de cette contrainte, la méthode lasso élimine certains paramètres.

Régression logistique

Y est une variable qualitative à valeurs dans $\{0,1\}$. On note $\widetilde{X} = (X_1, \ldots, X_p)$. La régression logistique se base sur l'hypothèse suivante :

$$\log \frac{\mathbb{P}(Y=1|\widetilde{X}=x)}{1-\mathbb{P}(Y=1|\widetilde{X}=x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Cela implique donc

$$\mathbb{P}(Y=1|\widetilde{X}=x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} = \frac{1+\tanh\left(\frac{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}{2}\right)}{2}.$$

Soit $\delta \in \{0,1\}$. Alors

$$\mathbb{P}(Y=\delta|\widetilde{X}=x) = \mathbb{P}(Y=1|\widetilde{X}=x)^{\delta} \times \mathbb{P}(Y=0|\widetilde{X}=x)^{1-\delta}.$$

Cela permet d'en déduire la vraisemblance L :

$$L(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y=y_i|\widetilde{X}=x) = \prod_{i=1}^{n} \mathbb{P}(Y=1|\widetilde{X}=x)^{y_i} \times \mathbb{P}(Y=0|\widetilde{X}=x)^{1-y_i}.$$

En passant au logarithme, on obtient :

$$\log L(\beta) = \sum_{i=1}^{n} \Big(y_i X_i.\beta - \log\left(1+e^{X_i.\beta}\right)\Big).$$

On note par $\hat{\beta}$ l'estimateur du maximum de la vraisemblance. On a alors le résultat suivant qui nous sera utile pour la suite :

Proposition 1. Pour tout $1 \leqslant i \leqslant p$, l'estimateur $\hat{\beta}_i$ converge en loi vers une loi normale de moyenne β_i et de variance $\mathrm{Var}(\hat{\beta}_i)$ lorsque n tend vers l'infini.

Logistique avec sélection de variables par le lasso

Ce modèle consiste à déterminer

$$\underset{\beta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \Big(y_i X_i.\beta - \log\left(1+e^{X_i.\beta}\right)\Big) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Cela revient à maximiser la vraisemblance pour la régression logistique tout en éliminant certaines variables avec la pénalisation lasso. Tout comme le lasso, le paramètre λ que l'on choisit est celui pour lequel le AIC obtenu est le plus faible.

Implémentation de la méthode sous R

La méthode logistique avec sélection des variables par le lasso a été implémentée par Vivian Viallon sur R. Elle requiert l'utilisation du package « glmnet ».

6.2.2 Forêts aléatoires

Bagging

On pose $Z = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ l'échantillon. Soit $\hat{f}(x)$ un modèle de prédiction de y fonction de x. Pour réduire la variance du modèle, on génère B échantillons bootstrap $Z^{*b}, b = 1, 2, \ldots, B$. Soit $\hat{f}^{*b}(x)$ la prédiction du modèle pour l'échantillon Z^{*b}.
Le bagging consiste alors à prendre comme modèle de prédiction la fonction [7]

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x).$$

Ce calcul induit l'erreur *out-of-bag*, notée Err_{oob}, ce qui permet de remplacer la validation croisée. Elle se calcule de la manière suivante :

$$\text{Err}_{\text{oob}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} L(y_i, \hat{f}^{*b}(x_i)) \tag{1}$$

où C^{-i} est l'ensemble des échantillons bootstrap ne contenant pas l'observation i, $|C^{-i}|$ le nombre de ces échantillons et L est la fonction de perte utilisée pour quantifier l'erreur de prédiction.

Forêts aléatoires

L'algorithme des forêts aléatoires est une variante du *bagging*. Si on dispose de B variables i.i.d., notées $Z_1, \ldots, Z_B$, de variance σ^2, alors

$$V\left(\frac{Z_1 + \cdots + Z_B}{B}\right) = \frac{\sigma^2}{B}.$$

Maintenant si les variables sont identiquement distribuées mais avec une corrélation ρ des variables prises deux à deux, alors

$$\begin{aligned} V\left(\frac{Z_1 + \cdots + Z_B}{B}\right) &= \frac{1}{B^2} \sum_{1 \leqslant i,j \leqslant B} \text{Cov}(Z_i, Z_j), \\ &= \frac{1}{B^2} \left(B\sigma^2 + 2 \sum_{1 \leqslant i < j \leqslant B} \text{Cov}(Z_i, Z_j) \right), \\ &= \frac{1}{B^2} \left(B\sigma^2 + B(B-1)\rho\sigma^2 \right), \\ &= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \end{aligned}$$

Si on augmente B, le terme $\frac{1-\rho}{B}\sigma^2$ diminue mais $\rho\sigma^2$ reste inchangé. Ainsi le *bagging* perd de son intérêt si la corrélation ρ est élevée. L'algorithme des forêts aléatoires cherche à rectifier ce problème en diminuant ρ. Pour cela, on retient au hasard m variables explicatives parmi les p disponibles, et ce pour chaque arbre de classification. Cela a pour effet de diminuer les corrélations entre les différents arbres. On combine les B arbres obtenus pour obtenir la fonction de prédiction. Plus précisément, l'algorithme des forêts aléatoires est le suivant :

Algorithme 1. Forêts aléatoires [7]

1. Pour b allant de 1 à B :
   (a) Générer un échantillon boostrap Z^{*b}.
   (b) Construire un arbre T_b sur l'échantillon Z^{*b}, en répétant les instructions suivantes pour chaque nœud externe de l'arbre, jusqu'à ce que le nombre minimal d'observations dans un nœud soit atteint.
      i. Choisir m variables au hasard parmi les p.
      ii. Choisir les variables les plus pertinentes parmi les m pour effectuer la classification.
      iii. Scinder le nœud en deux nœuds fils.
2. Retourner l'ensemble des arbres $\{T_b\}_1^B$.

Pour faire une prédiction à une nouvelle observation x :

Régression : $\hat{f}_{\text{rf}}^B(x) = \dfrac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification : Soit $\hat{C}_b(x)$ la classe prédite par l'arbre T_b. Alors $\hat{C}_{\text{rf}}^B(x) = \textit{vote majoritaire}\ \{\hat{C}_b(x)\}_1^B$.

En général on choisit $m = \sqrt{p}$ si y est une variable qualitative, $m = \dfrac{p}{3}$ si y est quantitative [7]. Mais rien n'empêche de choisir d'autres valeurs comme ci-dessous.

Cet algorithme permet également de déterminer la probabilité d'appartenir à une classe si la variable à expliquer est qualitative. Supposons que la réponse Y soit une variable qualitative. Soit $\{1, \ldots, k\}$ l'ensemble des valeurs de Y. Alors pour tout $1 \leqslant i \leqslant k$:

$$\mathbb{P}(\hat{Y} = i|x) = \frac{|\{1 \leqslant b \leqslant B \ / \ \hat{C}_b(x) = i\}|}{B}.$$

Choix des paramètres

Ainsi il y a deux paramètres que l'on peut modifier dans l'algorithme des forêts aléatoires, le nombre d'arbres B et le nombre de variables sélectionnées m. Cet algorithme étant basé sur le *bagging*, l'erreur *out-of-bag* (1) remplace la nécessité d'effectuer une validation croisée. Afin de choisir les paramètres, on a donc calculé cette erreur pour chaque couple de paramètres et on a retenu ceux qui donnent l'erreur *out-of-bag* la plus faible.
Les paramètres testés sont les suivants :

— $B = 500, 750, 1000$

— $m = 5, 10, 15, 20$

Sur nos données, l'erreur la plus faible a été obtenue pour :

$$B = 500, m = 5.$$

Implémentation de la méthode sous R

L'algorithme des forêts aléatoires a été implémenté en utilisant la fonction *randomForest* du package « randomForest ». En plus de donner la prédiction de la responsabilité « expert », il indique la probabilité prédite d'être responsable.

6.2.3 Boosting

Le boosting est un algorithme d'apprentissage machine. Le principe consiste à améliorer les performances d'un « classifieur faible » [7]. Pour ce faire, le *boosting* crée un nouveau modèle comme une version adaptative du précédent en donnant plus de poids aux observations mal prédites. Il a été originalement créé pour la prévision d'une variable binaire mais cet algorithme a été adapté à de nombreuses situations. On présente ici l'algorithme original appelé *AdaBoost* qui est approprié à l'étude étant donné que la responsabilité est une variable binaire.

L'algorithme *AdaBoost* est basé sur une variable à expliquer binaire à valeurs dans $\{-1, 1\}$, mais on l'adapte facilement pour n'importe quelle variable binaire à l'aide d'un changement de variable affine. Il construit des classifieurs $G_m(x)$, $1 \leqslant m \leqslant M$, et les combine en un classifieur final $G(x)$. L'algorithme est le suivant :

Algorithme 2. AdaBoost [7]
Soit x à prévoir.

1. Initialiser les poids $\omega_i = 1/n$, $1 \leqslant i \leqslant n$.
2. Pour m allant de 1 à M :
   (a) Créer un classifieur $G_m(x)$ à valeurs dans $\{-1, 1\}$ sur l'échantillon en utilisant les poids ω_i.
   (b) On calcule
   $$\mathrm{err}_m = \frac{\sum_{i=1}^{n} \omega_i |y_i - G_m(x_i)|}{\sum_{i=1}^{n} \omega_i}.$$
   (c) Calculer $\alpha_m = \log \dfrac{1 - \mathrm{err}_m}{\mathrm{err}_m}$.
   (d) On change les poids : $\omega_i \leftarrow \omega_i \cdot e^{\alpha_m |y_i - G_m(x_i)|}$, $\quad 1 \leqslant i \leqslant N$.
3. Retourner $G(x) = \mathrm{sgn}\left(\sum_{m=1}^{M} \alpha_m G_m(x) \right)$.

Remarque 1. Dans notre étude, la variable de responsabilité est à valeurs dans $\{0, 1\}$, on peut la ramener à $\{-1, 1\}$ en posant $\widetilde{y}_i = 2y_i - 1$, et prendre comme classifieur final $\widetilde{G}(x) = \dfrac{G(x) + 1}{2}$.
On peut montrer que l'algorithme 2 est équivalent à l'algorithme suivant :

Algorithme 3. Forward Stagewise Additive Modeling [7]
On suppose ici que y est à valeurs dans $\{-1, 1\}$.

1. Initialiser $f_0(x) = 0$.
2. Pour m allant de 1 à M :
   (a) Déterminer
   $$(\beta_m, G_m) = \arg\min_{\beta, G} \left\{ e^{-y_i \left(f_{m-1}(x_i) + \beta G(x_i) \right)} \right\}.$$
   (b) Poser $f_m(x) = f_{m-1}(x) + \beta_m G_m(x)$.
3. Retourner $\mathrm{sgn}\big(f_M(x)\big)$.

M représente le nombre d'arbres ou de classifieurs à créer. Chaque classifieur G_m peut être un arbre de classification, dont la profondeur est à fixer. On peut également utiliser un paramètre de *shrinkage* $0 < s \leqslant 1$ de la manière suivante dans l'algorithme :

$$f_m(x) = f_{m-1}(x) + s \cdot \beta_m G_m(x).$$

Insérer un paramètre de *shrinkage* permet de limiter les risques de surapprentissage, mais en contrepartie il faut plus d'itérations pour avoir un bon résultat.

L'algorithme du *boosting* permet également de donner la probabilité pour la prédiction d'être dans telle classe.

Proposition 2. En reprenant les notations de l'algorithme 3, si Y est la variable binaire à expliquer, on a [7] :

$$\mathbb{P}(\hat{Y} = 1|x) = \frac{1 + \tanh\big(f_M(x)\big)}{2}.$$

Cela justifie au passage pourquoi on prend $\operatorname{sgn}\big(f_M(x)\big)$ dans l'algorithme 3 pour déterminer la classe de la prédiction.

Choix des paramètres

Les paramètres que l'on décide d'optimiser dans l'algorithme sont le nombre d'arbre M, la profondeur de chaque arbre h et le paramètre de *shrinkage* s. On effectue sur l'échantillon Z une méthode de validation croisée avec $k = 5$, pour déterminer les paramètres pour lesquels la moyenne des erreurs sur les échantillons tests, est la plus faible.
Les paramètres testés sont les suivants :

- $M = 500, 1000, 1500, 2000$
- $h = 4, 5, 6, 7, 8$
- $s = 10^{-3}, 2 \times 10^{-3}, 3 \times 10^{-3}, 4 \times 10^{-3}, 5 \times 10^{-3}$

Sur nos données, l'erreur la plus faible a été obtenue pour :

$$M = 2000, h = 6, s = 5 \times 10^{-3}.$$

Implémentation de la méthode sous R

La méthode du *boosting* a été mise en place sur R par la fonction *gbm* du package « gbm ». Il donne également les probabilités pour chaque usager d'être responsable.

6.3 Courbe ROC et AUC

La courbe ROC est une méthode utilisée en statistiques pour mesurer le pouvoir discriminant d'un classificateur à valeurs continues, avec la réponse appartenant à deux classes. On dispose d'observations et d'une fonction de prédiction $\hat{f}$. Quitte à faire un changement de variables affines, on suppose que la réponse (c'est-à-dire y) doit être à valeurs dans $\{-1, 1\}$.

Soit $\hat{y}_i = \hat{f}(x_i) \in \mathbb{R}$. Pour déterminer à quelle classe la réponse $\hat{y}_i$ appartient, on définit un seuil $s \in \mathbb{R}$ tel que

$$\hat{y}_i = \begin{cases} 1 \text{ si } \hat{y}_i \geqslant s, \\ -1 \text{ sinon.} \end{cases}$$

On définit la classe positive comme celle où $y = 1$ et la classe négative comme celle où $y = -1$. La matrice de confusion est la suivante :

Classe réelle \ Classe prédite	Positive	Négative
Positive	VP (Vrai Positif)	FN (Faux Négatif)
Négative	FP (Faux Positif)	VN (Vrai Négatif)

TABLE 6 – Matrice de confusion

La sensibilité est le taux de vrai positif, noté TVP. On a $\text{TVP} = \dfrac{\text{VP}}{\text{VP} + \text{FN}}$.

La spécificité est le taux de vrai négatif, noté TVN. On a $\text{TVN} = \dfrac{\text{VN}}{\text{VN} + \text{FP}}$. On remarque que $1 - \text{TVN}$ est le taux de faux positif, noté TFP.

TVP et TVN dépendent du seuil s choisi. On calcule alors pour tout s, $\text{TVP}(s)$ et $\text{TFP}(s)$. Il suffit de restreindre s à $[\min_{1 \leqslant i \leqslant n} \hat{y}_i, \max_{1 \leqslant i \leqslant n} \hat{y}_i + 1]$, en effet :

$$\text{si } s \leqslant \min_{1 \leqslant i \leqslant n} \hat{y}_i, (\text{TVP}, \text{TFP}) = (1, 1),$$

$$\text{si } s > \max_{1 \leqslant i \leqslant n} \hat{y}_i, (\text{TVP}, \text{TFP}) = (0, 0).$$

On trace alors TVP en fonction de TFP. L'aire sous la courbe est notée AUC. Elle mesure le pouvoir discriminant du prédicteur, plus AUC est proche de 1, meilleur est le classificateur. Cette aire peut se calculer de manière approchée par la méthode des trapèzes, mais elle peut être déterminée de manière exacte.

6.4 Variables explicatives utilisées

Les variables utilisées pour l'étude sont les suivantes :

- obstacle_fixe_heurte : indique si le véhicule de l'usager a heurté un obstacle fixe
  0 - Aucun ou non renseigné
  1 - Oui
- obstacle_mobile_heurte : indique si le véhicule de l'usager a heurté un obstacle mobile
  0 - Aucun ou non renseigné
  1 - Oui
- point_choc : endroit ou le véhicule a été heurté
  0 - Non renseigné
  1 - Avant
  2 - Gauche
  3 - Droite
  4 - Arrière
- situation_risquee : le véhicule était entre deux files, effectuait un demi-tour ou était en marche arrière
  0 - Non
  1 - Oui
- insertion : Le conducteur s'insérait sur la voie
  0 - Non
  1 - Oui

- change_file : L'usager changeait de file
  0 - Pas de changement de file
  1 - Oui
- deporte : deportation du véhicule
  0 - Non
  1 - Oui
- deporte_gauche : deportation du véhicule à gauche
  0 - Non
  1 - Oui
- tourne : véhicule tournant dans une direction
  0 - Non
  1 - Oui
- tourne_gauche : véhicule tournant à gauche
  0 - Non
  1 - Oui
- depasse : le véhicule dépassait un autre
  0 - Non
  1 - Oui
- depasse_gauche : le véhicule dépassait un autre à gauche
  0 - Non
  1 - Oui
- evitement : l'usager effectuait une manoeuvre pour eviter l'accident
  0 - Non
  1 - Oui
- immobile : le véhicule était arrêté, en stationnement (avec occupants) ou effectuait une manoeuvre de stationement
  0 - Non
  1 - Oui
- cadm : catégorie administrative de la route sur laquelle s'est produit l'accident
  1 - Autoroute
  2 - Route nationale
  3 - Route départementale
  9 - Autre
- regime : régime de circulation
  0 - Non renseigné ou sans objet
  1 - A sens unique
  2 - Bidirectionnelle ou avec voies d'affectation variable
  3 - A chaussées séparées
- nbvoies : nombre total de voies de circulation
  0 - Non renseigné, route à cinq voies, autre ou inconnu
  1 - Route à sens unique
  2 - Route bidirectionnelle
  3 - Route bidirectionnelle à trois voies
  4 - Route à chaussées séparées à 2×2 voies ou route bidirectionnelle à quatre voies
- pente : la route était en pente
  0 - Non
  1 - Oui

- courbe : la route était courbée
  0 - Non
  1 - Oui
- terre_plein_central : présence d'un terre-plein central sur le lieu de l'accident
  0 - Non
  1 - Oui
- largrout : largeur de la route
- surface : état de la surface de la route
  0 - Normale
  1 - Mouillée, boue, corps gras, verglacé, enneigé ou inondée ou autre
- amenag : aménagement de la route
  0 - Non renseigné ou sans objet
  1 - Souterrain, tunnel, pont ou autopont
  2 - Bretelle d'échangeur ou de raccordement
  3 - Carrefour aménagé
- chaussee : indique si l'accident a eu lieu sur la chaussée
  0 - Oui
  1 - Ailleurs
- lumiere : luminosité du lieu au moment de l'accident
  0 - Plein jour
  1 - Aube, crépuscule ou nuit mais avec éclairages
  2 - Nuit sans éclairage
- local : indique si l'accident a eu lieu en agglomération
  1 - Hors agglomération
  2 - En agglomération
- int : type d'intersection où l'accident a eu lieu
  0 - Hors intersection
  1 - En X (4 branches)
  2 - En T ou Y (3 branches)
  3 - Giratoire
  9 - Autre
- climat_anormal : climat non doux au moment de l'accident
  0 - Non
  1 - Oui
- forte_pluie : forte pluie au moment de l'accident
  0 - Non
  1 - Oui
- pire_climat : brouillard, neige ou tempête au moment de l'accident
  0 - Non
  1 - Oui
- changement_direction_surprise : le véhicule effectuait un changement de direction sans avertissement préalable
  0 - Non
  1 - Oui
- comportement_dangereux : Imprudence délibérée, depassement dangereux ou sens interdit de la part du conducteur
  0 - Non
  1 - Oui

- mauvaise_voie : Le véhicule était sur une voie interdite pour lui
  0 - Non
  1 - Oui
- refus_priorite : l'usager a refusé une priorité
  0 - Non
  1 - Oui
- vitesse_excessive : le véhicule dépassait la vitesse maximale autorisée
  0 - Non
  1 - Oui
- nombre_fautes : indique le nombres de fautes commises par l'usager. Les fautes correspondent à avoir une valeur égale à 1 pour les variables suivantes :
  - changement_direction_surprise
  - comportement_dangereux
  - mauvaise_voie
  - refus_priorite
  - situation_risquee
  - vitesse_excessive
  - Arrêt, stationnement gênant ou interdit
  - Distance de sécurité non respectée
  - Feux non allumés en situation de nécessité
  - Vitesse excessivement lente du véhicule

  Remarque : cela revient à sommer toutes ces variables

Références

[1] S. Blaizot, F. Papon, M. M. Haddak, and E. Amoros. Injury incidence rates of cyclists compared to pedestrians, car occupants and powered two-wheeler riders, using a medical registry and mobility data, rhone county, france. 58 :35–45.

[2] L Bouaoun, M. Haddak, and A. Amoros. Fatal road traffic crashes : comparisons by road user types and measures of exposure. 75 :217–225.

[3] Jeff Brubacher, Herbert Chan, and Mark Asbridge. Development and Validation of a Crash Culpability Scoring Tool. *Traffic Injury Prevention*, 13(3) :219–229, May 2012.

[4] A.E. af Wåhlberg and L. Dorn. Culpable versus non-culpable traffic accidents ; What is wrong with this picture ? *Journal of Safety Research*, 38(4) :453–459, January 2007.

[5] Jeff Brubacher, Herbert Chan, and Mark Asbridge. Culpability analysis is still a valuable technique. *International Journal of Epidemiology*, 43(1) :270–272, 2014.

[6] Susantha Chandraratna and Nikiforos Stamatiadis. Quasi-induced exposure method : Evaluation of not-at-fault assumption. *Accident Analysis & Prevention*, 41(2) :308–313, March 2009.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer. Second edition edition, 2011.

[8] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013. DOI : 10.1007/978-1-4614-7138-7.

[9] Edouard Ollier and Vivian Viallon. Regression modelling on stratified data with the lasso. *Biometrika*, 104(1) :83–96, 2017.

[10] Michael D. Robertson and Olaf H. Drummer. Responsibility analysis, a methodology to study the effects of drugs in driving. 26 :243–247, 1994.