EURIA
Euro-Institut d'Actuariat

siapartners

Groupe 6 :

Tarification d'un produit IARD et analyses d'impacts sur un marché concurrentiel

Alexandre BOUTEMY
Antoine GODEC
Mathilde MOUDEN

Dirigé par :
Romain LAILY (Actuaire)
Wassim YOUSSEF (Actuaire)
Franck VERMET (Enseignant-Chercheur)

En partenariat avec le cabinet de conseil SIA PARTNERS

Bureau d'étude 2015-2016
Master 1 Actuariat
Euro-Institut d'Actuariat

Remerciements

Nous souhaitons, en préliminaire à ce rapport, faire part de notre gratitude aux personnes qui ont contribué à sa réalisation.

Nous tenons à remercier chaleureusement Messieurs Wassim YOUSSEF et Romain LAILY, actuaires au cabinet de conseil SIA PARTNERS, pour nous avoir proposé ce sujet de bureau d'études particulièrement intéressant. Nous leur savons gré de leurs précieux conseils face à nos interrogations, de leur disponibilité et de de leur promptitude à nous répondre.

Nous tenons aussi à exprimer notre reconnaissance à Monsieur Franck VERMET, enseignant-chercheur à l'EURIA et tuteur de notre bureau d'étude. Son écoute, ses critiques et son suivi tout au long du projet ont permis à celui-ci de progresser.

Enfin, nous adressons nos vifs remerciements à Monsieur Pierre AILLIOT pour l'aide précieuse qu'il nous a apportée, notamment en informatique.

Table des matières

Liste des tableaux

Table des figures

Introduction

L'enjeu majeur auquel est confronté toute compagnie d'assurance est la bonne évaluation des risques auxquels elle doit faire face ; elle se doit par ailleurs de conserver ou d'accroitre sa part de marché dans un secteur concurrentiel.

Des modèles statistiques y sont de ce fait utilisés quotidiennement, permettant la prise en compte d'un certain nombre de données et de variables. L'incroyable avancée technologique (et notamment informatique) des années 2000 se traduit par des capacités de calculs toujours supérieures, ce qui rend possible l'exploitation de gigantesques bases de données : le Big Data. Les assureurs voient en ces mégadonnées l'opportunité de se démarquer face aux concurrents : si un nombre considérable de variables peut maintenant être traité, cela permet une meilleure connaissance de l'assuré qui est prêt à dévoiler le plus d'informations personnelles possible en échange d'une réduction des tarifs qui lui sont proposés. Bien que les réflexions sur le Big Data ne font que commencer, les grandes compagnies n'hésitent pas à se lancer dans l'aventure, notamment avec le concept du Pay As You Drive [1].

C'est dans ce contexte que le cabinet de conseil SIA PARTNERS a proposé à l'EURIA une problématique de bureau d'étude intitulée « Tarification d'un produit d'assurance IARD et analyses d'impacts sur un marché concurrentiel ».

Afin d'initier les élèves de Master 1 au métier d'actuaire, l'EURIA propose en effet aux étudiants des thèmes de bureau d'étude provenant d'entreprises partenaires. Les sujets concrets permettent d'appliquer une grande partie des notions vues au cours des deux premières années de la formation. Les étudiants sont encadrés par un ou plusieurs tuteur(s) universitaire(s) et tuteur(s) professionnel(s). Ils ont ainsi l'opportunité de se familiariser avec le monde professionnel et de tisser des liens privilégiés avec l'entreprise partenaire.

Ce bureau d'étude a été encadré par Franck VERMET, enseignant-chercheur à l'Université de Bretagne Occidentale, par Wassim YOUSSEF et Romain LAILY tous deux actuaires au cabinet de conseil SIA PARTNERS. Bien que nos travaux n'aient pas un impact direct sur les méthodes de tarification de SIA PARTNERS, le cabinet s'y intéresse. En effet, comme évoqué précédemment, les techniques d'actuariat sont en constante évolution, et une entreprise de cette ampleur s'en informe régulièrement afin de rester compétitive.

Il s'agit d'un bureau d'étude particulier puisqu'il est issu d'un jeu organisé durant l'été 2015 par Arthur CHARPENTIER, membre de l'Institut des Actuaires. L'objectif du jeu est d'identifier les modèles informatiques qui prédisent au mieux la prime pure individuelle (ou tarif technique) [2] dans les contrats d'assurance automobile. La prédiction de la prime est correcte si la somme des primes de tous les assurés couvre le montant total des sinistres. Cette couverture peut être représentée à l'aide du ratio S/P.

Le jeu s'est déroulé sous forme d'un concours, auquel de nombreux actuaires ont participé. A l'issue de celui-ci, l'institut des actuaires a organisé un colloque SCOR/Institut des actuaires « Recherche actuarielle et Data science » afin de dévoiler les résultats aux participants du concours.

Ce sujet nous a particulièrement intéressés puisqu'il constitue un exercice typique du métier d'actuaire : la tarification de contrats automobiles. Nous souhaitions par ailleurs nous confronter à un sujet d'actualité actuarielle : l'évolution des méthodes et des outils d'actuariat.

1. Installation d'un boitier à l'intérieur du véhicule pour évaluer les habitudes de conduite d'un assuré. Une conduite prudente réduit le tarif.
2. Le tarif réel payé par l'assuré est composé d'un tarif technique qui correspond à ce que l'assuré coûtera probablement à l'assureur et d'une marge commerciale.

Nous avons pris ce bureau d'étude comme un challenge, puisqu'il nous permet de nous mesurer, à notre échelle, à des actuaires expérimentés.

Franck VERMET ayant participé au colloque SCOR/Institut des actuaires, nous avons pu discuter ensemble des résultats. Cela nous a permis de nous approprier l'objectif du projet : identifier les modèles informatiques les plus performants dans la prédiction des montants des sinistres. Un bon modèle est un modèle qui permet à l'assureur de financer la totalité des indemnités versées aux assurés grâce aux primes perçues. L'indicateur de la performance des modèles est donc le ratio S/P.

L'objectif ainsi défini nous a amené, en endossant le rôle d'un assureur en date du 1er Janvier 2011, à décomposer notre travail en deux parties (après un traitement initial des données).

La première partie vise à la prédiction du nombre et du coût des sinistres des assurés en utilisant différents modèles de tarification, ceci grâce à deux bases de données mises à disposition par Arthur CHARPENTIER lors du concours :

- `Training` qui contient les caractéristiques de 100 021 anciens assurés ainsi que le montant de leurs sinistres pour les années 2009 et 2010.
- `Pricing` qui contient les caractéristiques de 36 311 clients potentiels (différents des assurés observés dans la base `Training`) à assurer en 2011. Le montant de leur sinistre en 2011 n'est pas connu (c'est une variable constatée au 31 Décembre 2011). Nous devons en tant qu'assureur estimer ce montant pour chaque assuré à l'aide d'un modèle informatique. L'estimation obtenue correspond à la prime demandée au client au 1er Janvier 2011, en échange de l'assurance.

Parmi les nombreux modèles de tarification susceptibles d'être utilisés, 9 ont été retenus. Chaque modèle est associé à un assureur. A la fin de cette phase, nous obtenons pour chacun des 36 311 assurés, 9 primes distinctes : chaque assureur propose ainsi un prix à tous les assurés.

La deuxième partie consiste d'abord à simuler un marché mettant en concurrence les 9 assureurs. Chaque assureur propose son tarif aux 36 311 clients potentiels, qui choisissent alors de s'assurer :

- Dans un premier temps, chez l'assureur le moins cher.
- Dans un second temps, au hasard chez l'un des assureurs les moins chers.

Une fois le marché simulé, les assurés sont répartis entre les différents assureurs, chacun disposant ainsi de son propre portefeuille.

Ensuite, il s'agit de calculer le ratio S/P pour chaque assureur. L'activité de l'assureur est viable si son ratio S/P est inférieur ou égal à 1. Les assureurs ayant obtenant un ratio S/P inférieur ou égal à 1 sont les assureurs qui utilisent un bon modèle. L'assureur qui obtient le ratio S/P le plus faible est celui qui utilise le meilleur modèle.

Une telle démarche doit nous permettre d'évaluer la performance des différents modèles, donc de répondre à la problématique de ce bureau d'étude.

Chapitre 1

Considérations générales

Afin de bien comprendre la démarche, puis la réalisation du projet, il est utile de rappeler succinctement quelques considérations générales.

1.1 L'assurance automobile en France [1]

La tarification des contrats automobiles étant le support de notre projet, un bref panorama de l'assurance automobile en France s'impose.

Il s'agit du deuxième marché non-vie (derrière l'assurance Dommage Corporelle), représentant 21% du chiffre d'affaire total du marché de l'assurance non-vie, soit 16,1 milliards d'euros. On dénombre 40,2 millions de véhicules assurés en métropole répartis entre une centaine d'assureurs.

La seule garantie obligatoire de l'assurance automobile est la Responsabilité Civile (RC) ou assurance au tiers. Elle permet d'indemniser les dégâts matériels causés aux autres véhicules (RC matérielle) et d'indemniser les blessures ou décès causés à un tiers (piéton, passager, conducteur d'un autre véhicule) (RC corporelle).

D'autres garanties, optionnelles, peuvent être souscrites, telles que les dommages aux véhicules, le bris de glace, le vol, l'incendie, les catastrophes naturelles qui couvrent les dégâts subis par le véhicule de l'assuré ou encore les dommages corporels qui couvrent les blessures ou le décès de l'assuré.

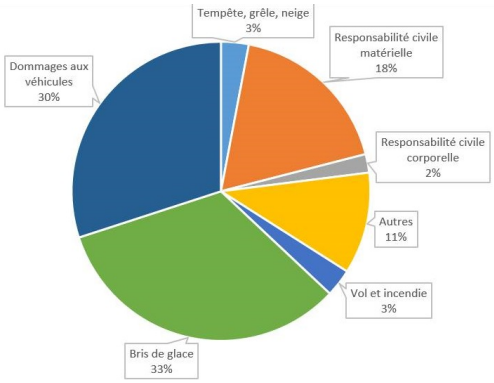

FIGURE 1.1 – *Répartition en pourcentage du nombre de sinistres en 2014*

On estime à 7,7 millions le nombre de sinistres automobiles déclarés en 2014. La Responsabilité Civile ne représente que 20% du nombre de sinistres (18% pour la RC matérielle et 2% pour la RC corporels, mais 49% des charges annuelles (16% pour la RC matériels et 33% pour la RC corporels). C'est donc une garantie qui coûte cher à l'assureur.

1. source FFSA, année 2014

Les indemnités sont financées grâce aux cotisations payées par les assurés. La prime annuelle moyenne pour l'assurance automobile est de 400 €, dont 175 € se rapportent à la garantie RC globale (matériels et corporels).

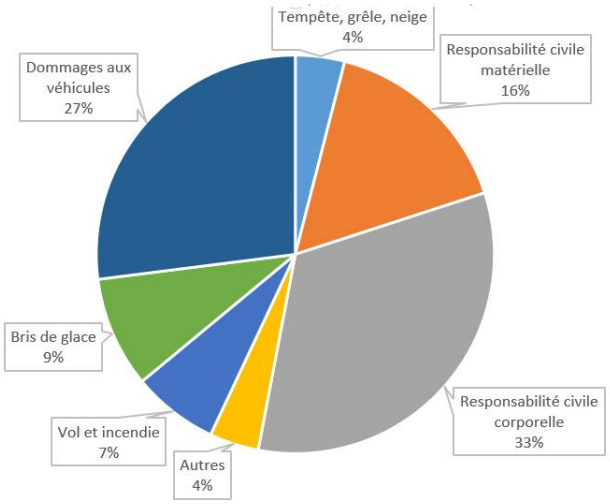

FIGURE 1.2 – *Répartition en pourcentage du montant des sinistres en 2014*

1.2 Le ratio S/P

Au travers de la tarification d'un produit d'assurance, l'objectif de tout assureur est d'assurer l'équilibre entre le paiement des indemnités versées à l'assuré suite aux sinistres constatés pendant l'année d'exercice et la cotisation versée par l'assuré en début d'exercice.

Afin de trouver le meilleur équilibre tarifaire, le ratio S/P est un indicateur utilisé par les assureurs. Il s'agit du rapport entre :

- S, le montant des sinistres constatés de l'assuré. En d'autres termes, il s'agit du montant des indemnités versées par l'assureur à l'assuré.
- P, le montant des primes récoltées par l'assureur.

Le montant de la prime réellement payée par l'assuré constitue la prime commerciale. Celle-ci se décompose en deux parties :

- La prime pure qui prend uniquement en compte le volet technique du coût de l'assuré, c'est-à-dire le nombre et le coût moyen des sinistres. On peut la représenter par l'équation suivante :

$$Prime\ pure = Nombre\ de\ sinistres \times Coût\ moyen\ d'un\ sinistre$$

- Une part commerciale, qui intègre essentiellement les frais de gestion, les taxes et la marge de l'assureur, et est impactée par la politique commerciale de l'assureur.

On peut aussi écrire plus simplement :

$$Prime\ commerciale = Prime\ pure + marge\ commerciale$$

Si le ratio S/P est égal à 90%, cela signifie que pour 100 € de primes perçues, l'assuré a coûté 90 € à l'assureur. On comprend alors que si le ratio inférieur ou égal à 1, les primes perçues couvrent entièrement les indemnités versées à l'assuré. A contrario, si ce ratio est supérieur à 1, les cotisations ne parviennent plus à couvrir les indemnisations.

Un ratio S/P individuel peut donc être calculé assuré par assuré. Mais cela ne constitue pas pour la compagnie d'assurance le meilleur indicateur de « bonne prédiction du tarif », puisque l'activité d'assurance

se base sur le principe de mutualisation du risque : les cotisations de tous les assurés, qu'ils aient eu un sinistre ou non, servent à financer les indemnités versées aux victimes.

Le ratio S/P global calculé comme le montant global des sinistres que divise le total des primes encaissées, permettra de mieux évaluer l'équilibre financier de la compagnie d'assurance. Ainsi, un ratio S/P global supérieur à 1 est l'indicateur d'une activité déficitaire. Il est donc nécessaire que la somme des primes de tous les assurés couvre le montant total des sinistres pour que l'activité soit rentable, ou tout au moins équilibrée. Un ratio S/P global équivalent à 1 indique que l'assureur a tarifé au plus juste les contrats de ses clients. Tandis qu'un ratio S/P inférieur à 1 permet à l'assureur de dégager un bénéfice.

1.3 La nécessité de la prédiction par des modèles informatiques

Dans une majorité d'activités, le prix de revient est connu avant la vente. Ce n'est pas le cas l'activité d'assurance où le cycle de production est inversé : le prix de revient n'est connu qu'après la vente. En effet, l'assureur réclame au 1er Janvier une prime P pour un service, sans en connaitre le coût qui n'est autre que le montant des indemnités S qu'il devra verser à son assuré pendant l'année à venir. S n'est donc connu qu'en fin d'exercice, c'est-à-dire le 31 Décembre.

Par conséquent, le ratio S/P n'étant réellement connu qu'en fin d'année, l'assureur doit anticiper ce montant d'indemnités. Cette prédiction se fait à l'aide de nombreux modèles informatiques, sur la base d'historiques ou de profils similaires. Il s'agit de prédire les variables à expliquer, à savoir dans un premier temps le nombre de sinistres par assuré et dans un deuxième temps le coût moyen d'un sinistre. En multipliant alors les deux prédictions, on obtient le montant des indemnités estimées. Ce montant correspondra de fait à la prime pure demandée à l'assuré.

1.4 Les bases de données

Afin de réaliser la prédiction à l'aide d'un modèle informatique, une première base de données est nécessaire. Elle est appelée base d'apprentissage. Elle contient un portefeuille d'assurés doté des variables explicatives, liées à l'assuré (âge, sexe...) et à son véhicule (type, valeur...) et de variables à expliquer liées à la sinistralité de l'assuré. Le logiciel analyse alors les relations entre les variables explicatives et les variables à expliquer.

La seconde base de données, dite base test, contient le portefeuille objet de la prédiction. Il s'agit de clients potentiels. Le logiciel utilise le lien établi entre la variable à expliquer et les variables explicatives sur la base d'apprentissage pour prédire les variables à expliquer de la base test, et donc l'estimation du montant des indemnités pour l'année à venir. La prédiction est réalisée en fonction du profil des assurés de la base test, par comparaison avec les assurés de la base d'apprentissage. Une prime pure peut alors être demandée à un assuré ou proposée à un client potentiel.

Les résultats obtenus seront différents en fonction du modèle informatique retenu.

Rappelons à ce sujet que l'objectif de notre bureau d'étude est en priorité d'identifier les modèles informatiques les plus performants dans la prédiction des montants des sinistres.

1.5 La simulation du marché

Si un bon modèle permet à l'assureur de financer la totalité des indemnités versées aux assurés grâce aux primes perçues, il se doit aussi d'être attractif vis-à-vis des assurés sur un marché concurrentiel. Notre projet vise donc aussi à analyser l'impact des modèles retenus dans ce marché concurrentiel.

Par souci de concision, la simulation de marché réalisée sera traitée ultérieurement.

Chapitre 2

Les outils et la démarche

2.1 Les bases de données à notre disposition : `Training` et `Pricing`

Dans le jeu-concours qu'il a organisé, Arthur CHARPENTIER a fourni deux bases de données aux participants : `Training` et `Pricing`.

- La base de données nommée `Training` sert de base d'apprentissage au logiciel. C'est une base de données complète, qui comporte toutes les caractéristiques des assurés ainsi que le nombre et le montant des sinistres déclarés. Elle comporte 100 021 observations (représentant chacune un assuré) réparties sur deux années : 50 021 observations pour l'année 2009 et 50 000 autres pour l'année 2010. Cependant, les assurés sont différents d'une année à l'autre (on retrouve uniquement 21 clients assurés à la fois en 2009 et 2010). Les observations ne sont pas considérées comme un historique : les clients assurés en 2009 n'ont pas été distingués de ceux assurés en 2010. Cette base constitue le portefeuille de l'assureur.
- La base de données nommée `Pricing` est la base test. Elle comporte 36 311 observations, c'est-à-dire 36 311 clients potentiels pour l'année 2011.

`Training` et `Pricing` ont 15 variables explicatives en commun :

- `PolNum` : le numéro de la police d'assurance (à chaque assuré correspond un numéro)
- `CalYear` : l'année d'assurance (2009 ou 2010).
- `Gender` : le sexe de l'assuré
- `Age` : l'âge de l'assuré
- `Occupation` : la profession de l'assuré (5 professions)
- `Bonus` : le bonus de l'assuré exprimé en pourcentage. Il va de -50 (bonus) à 150 (malus)
- `PolDur` : l'ancienneté d'assurance de l'assuré chez l'assureur
- `Type` : le type du véhicule (5 types)
- `Category` : la catégorie du véhicule (3 catégories)
- `Group1` : le groupe du véhicule (20 groupes)
- `Value` : la valeur du véhicule
- `Adind` : la présence d'une garantie matérielle (1 si l'assuré a souscrit une garantie matérielle, 0 si-non)
- `Group2` : la région du domicile de l'assuré (10 régions)
- `SubGroup2` : la sous-région du domicile de l'assuré (471 sous-régions)
- `Density` : la densité de la ville du domicile de l'assuré

`Training` contient 5 variables supplémentaires connues à la fin de l'année 2010 :

- `Exppdays` : le nombre de jours d'exposition du véhicule
- `Numtppd` : le nombre de sinistres matériels de l'assuré au cours de l'année
- `Numtpbi` : le nombre de sinistres corporels de l'assuré au cours de l'année
- `Indtppd` : le montant total des sinistres matériels de l'assuré
- `Indtpbi` : le montant total des sinistres corporels de l'assuré

Concernant la base de test `Pricing`, ces 5 variables supplémentaires n'ont pas été initialement fournies aux participants du concours, dans la mesure où il s'agissait des variables à prédire pour l'année 2011. Elles

PolNum	CalYear	Gender	Type	Category	Occupation	Age	Group1	Bonus	Poldur	Value	Adind	SubGroup2	Group2	Density	Exppdays	Numtppd	Numtpbi	Indtppd	Indtpbi
200114978	2009	Male	C	Large	Employed	25	18	90	3	15080	0	L46	L	72.01288	365	1	0	0.0000	0.00000
200114994	2009	Male	E	Large	Employed	20	11	30	2	22370	1	O38	O	39.55041	365	1	0	0.0000	0.00000
200115001	2009	Female	E	Large	Unemployed	42	11	150	0	39650	0	Q28	Q	169.52915	365	2	0	0.0000	0.00000
200115011	2009	Female	C	Medium	Housewife	21	5	0	0	12600	1	L6	L	58.89469	365	1	0	0.0000	0.00000
200115015	2009	Female	D	Large	Employed	33	12	30	10	9065	0	N4	N	109.63189	365	2	0	0.0000	0.00000
200115016	2009	Female	D	Small	Employed	26	13	40	7	27335	1	N16	N	47.98270	365	1	0	0.0000	0.00000
200115023	2009	Female	C	Small	Unemployed	20	7	80	13	7710	0	Q65	Q	77.73739	365	1	0	0.0000	0.00000
200115043	2009	Female	B	Medium	Employed	29	3	-20	12	8965	0	R19	R	272.96699	365	1	0	0.0000	0.00000
200115048	2009	Male	E	Medium	Unemployed	31	3	-40	10	21030	1	R9	R	251.43285	355	1	0	0.0000	0.00000
200115063	2009	Male	D	Medium	Employed	35	7	120	1	19995	1	Q22	Q	144.99890	365	1	0	0.0000	0.00000
200115068	2009	Male	C	Small	Employed	27	11	30	0	18395	0	Q13	Q	166.55494	365	1	0	0.0000	0.00000
200115070	2009	Female	A	Large	Housewife	65	9	-30	15	11880	0	R35	R	223.30857	365	1	0	0.0000	0.00000
200115074	2009	Female	A	Medium	Unemployed	25	9	20	2	23130	0	M1	M	107.81705	365	1	0	0.0000	0.00000

FIGURE 2.1 – *Présentation de la base* `Training`

ne leur ont été dévoilées que le jour du colloque.

Nous avons vécu lors de notre BE une situation identique, à la différence fondamentale près que nous n'avons pu obtenir les 5 variables supplémentaires pour la base `Pricing`.

2.2 Les modèles utilisés

La prédiction sera réalisée à l'aide de 9 modèles choisis. Beaucoup d'entre eux ont été étudiés à l'EURIA, comme les modèles de régression linéaire et les algorithmes d'apprentissage statistique. D'autres algorithmes ont été utilisés, tel que le modèle boosting, très performant selon Arthur CHARPENTIER, et ont donc nécessité un travail de recherche plus approfondi de notre part. Les différents modèles seront présentés en temps utile.

Les modélisations des deux variables à prédire, nombre et coût moyen des sinistres, seront effectuées séparément, éventuellement à l'aide de modèles différents.

Le logiciel *R* établira un lien entre la variable à expliquer et la variable explicative à l'aide des différents modèles retenus.

Plus généralement, le support informatique du projet est le logiciel *R*. Notre choix s'est naturellement porté sur ce logiciel puisque c'est celui que nous maîtrisons le mieux à ce jour. Nous y avons notamment importé les deux bases de données `Training` et `Pricing`.

2.3 Démarche

Notons en préambule que :

- N'ayant obtenu aucune indication sur la conception d'une prime commerciale, seule la prime pure sera retenue dans la suite de notre travail.
- Chacun des 9 modèles étudiés sera représenté par un assureur, les 9 assureurs étant mis en compétition dans la simulation de marché réalisée.

Rappelons aussi que notre bureau d'étude concerne uniquement la tarification de la Responsabilité Civile automobile.

La démarche adoptée peut être décomposée en plusieurs étapes.

L'étape préliminaire consiste en un traitement des données fournies par Arthur CHARPENTIER. Ce travail indispensable permet de comprendre les données, d'éliminer les données inutiles, redondantes ou aberrantes qui pourraient fausser le calibrage d'un modèle, c'est-à-dire la relation entre la variables à expliquer et les variables explicatives - et donc biaiser la prédiction des primes -, et enfin de dresser le profil général du portefeuille.

Une première étape dite de prédiction va permettre de prédire la prime pure pour l'année 2011 de chacun des 36 311 assurés potentiels de la base `Pricing`. Le nombre des sinistres et le coût moyen d'un sinistre pour chaque assuré seront donc les variables à expliquer.

Leur modélisation se fera sur la base d'apprentissage `Training`. Le logiciel R établira le lien entre les variables à expliquer et les variables explicatives, à l'aide des 9 modèles retenus.

Chacun des 9 modèles sera projeté sur la base test `Pricing` dans le but de prédire à partir des caractéristiques des assurés constituant la base test, le nombre des sinistres d'une part et le coût moyen d'un sinistre d'autre part pour l'année 2011. Comme expliqué précédemment, la prime pure de chaque assuré sera obtenue en multipliant le nombre estimé des sinistres par le coût moyen d'un sinistre.

A la fin de cette phase, à chacun des 36 311 assurés sont associées 9 primes individuelles : chaque assureur (chaque modèle) est alors en capacité de proposer un prix (représenté par la prime pure) à tous les assurés.

La deuxième étape du projet, débute par une simulation et visera à simuler un marché au 1er Janvier 2011, mettant en concurrence les 9 assureurs (les 9 modèles de prédiction testés précédemment). Sur ce marché, un assuré aura donc connaissance de toutes les primes pures qui lui seront proposées, qui correspondront aux primes qu'il paiera effectivement en cas de souscription de contrat.

Les assurés choisiront leur assureur de deux façons. Pour cela, deux types de marché seront simulés : un premier où ils choisiront l'assureur le moins cher et un deuxième où ils choisiront au hasard l'un des assureurs les moins chers. Ce dernier marché est plus réaliste, puisque les clients n'expriment pas nécessairement une préférence pour l'assureur le moins cher, bien que cela soit le critère dominant. Un jeune peut ainsi décider de s'assurer chez l'assureur de ses parents, pour bénéficier d'une partie de leur bonus.

A l'issue de cette simulation, les 36 311 assurés de la base `Pricing` seront alors répartis entre les différents assureurs présents sur le marché. Chaque assureur se trouvera alors affecté d'un portefeuille de nouveaux d'assurés pour l'année 2011.

Suite à la simulation, la deuxième étape se poursuivra en comparant les 9 assureurs, donc les 9 modèles de tarification. Nous nous projetterons à la date du 31 Décembre 2011 et constaterons le montant des indemnités versées à chaque assuré au cours de l'année écoulée. A cette fin, les 5 variables manquantes à la base `Pricing` seront récupérées auprès d'Arthur CHARPENTIER [1].

Le calcul du ratio S/P global pour chaque assureur sera à ce moment possible et nous pourrons répondre à la problématique du projet, qui est, rappelons-le une dernière fois, de repérer parmi les des 9 modèles étudiés, celui qui est le plus performant, à savoir celui qui présente le meilleur ratio S/P global. Le rapport se clôtura par une analyse de la relation entre une bonne prédiction du modèle et un bon ratio S/P.

La Figure 2.2 synthétise notre démarche.

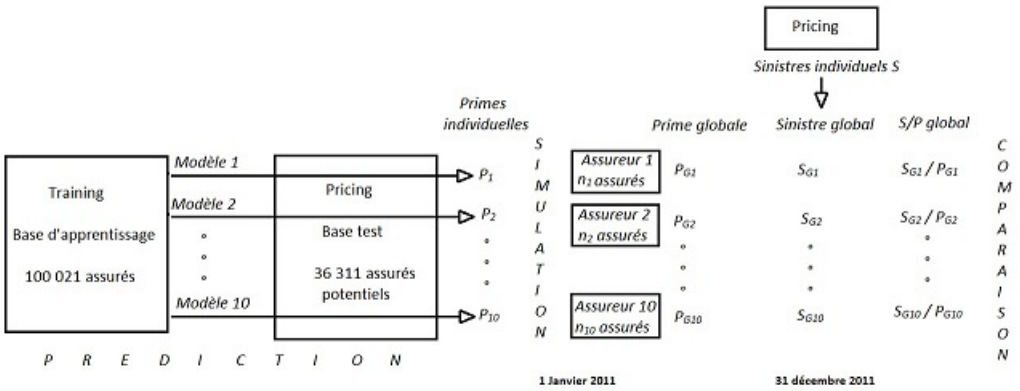

FIGURE 2.2 – *Schéma de la démarche suivie*

1. Nous ne sommes malheureusement pas parvenus à accéder à ces informations, ce qui nous a obligés à modifier notre stratégie. Cela sera développé au moment opportun dans la suite de ce rapport.

Chapitre 3

Traitement préalable des données

Cette étape est indispensable avant de débuter la prédiction. Elle permet en effet de comprendre les données disponibles, de repérer des valeurs aberrantes, de supprimer les variables redondantes ou inutiles, et de vérifier le format de ces variables.

Elle a été réalisée grâce aux 100 021 observations de la base `Training`, à l'aide du logiciel *R* et au moyen de grandeurs statistiques simples : moyenne, quantile, minium, maximum.

La répartition de chacune des variables au sein du portefeuille a été étudiée et son impact sur le nombre et le montant des sinistres matériels et corporels analysée. Les graphiques obtenus représentent les nombres et montants des sinistres en fonction de la variable analysée.

3.1 Analyse du portefeuille

Une bonne connaissance du portefeuille permet d'anticiper certains résultats, d'être capable d'esprit critique et d'éviter les mauvaises interprétations.

Les deux bases de données se présentent sous la forme de tableaux et contiennent deux types de variables :

- Les variables connues a priori par l'assureur, celles qu'il connait à la souscription du contrat.
- Les variables constatées a posteriori, celles que l'assureur constate à la fin de l'exercice.

Parmi les variables connues a priori, on distingue 3 catégories :

- Les variables liées à l'assuré : le numéro de la police d'assurance, l'âge, le sexe, la profession, le bonus, l'ancienneté d'assurance, la région d'habitation, la sous-région d'habitation, la densité et l'exposition.
- Les variables liées au véhicule : le type, la catégorie, le groupe et la valeur.
- La variable liée à la souscription de la garantie Responsabilité Civile matérielle.

La base de données `Training` servant de base d'apprentissage, c'est la seule qui contienne les variables constatées a posteriori. Il s'agit principalement de provisions [1].

Par ailleurs, le phénomène de sinistres « tardifs » [2] accentue le caractère approximatif des variables constatées a posteriori.

3.1.1 Variables liées à l'assuré

L'âge

A travers cette variable, l'assureur espère principalement récupérer une information sur l'expérience de l'assuré. Ces derniers ont un âge compris entre 18 et 75 ans. La tranche d'âge la plus présente au sein du

1. lorsque l'année est terminée, de nombreux sinistres corporels n'ont pas été totalement indemnisés : pronostic médical non établi, état instable. Le montant du sinistre corporel n'est alors qu'une estimation. En effet, si l'indemnisation des dommages matériels est quantifiable et rapide, l'indemnisation des sinistres corporels est incertaine et peut s'étendre sur plusieurs années.
2. sinistres survenus au cours de l'exercice mais déclarés après la clôture de celui-ci. Le nombre et le montant de ces sinistres doit être estimé.

portefeuille est celle des 20-50 ans. Le nombre de sinistres des 18-30 ans est plus de deux fois supérieur à celui des plus de 30 ans.

Le même constat peut être fait quant à l'impact de l'âge sur le coût des sinistres.

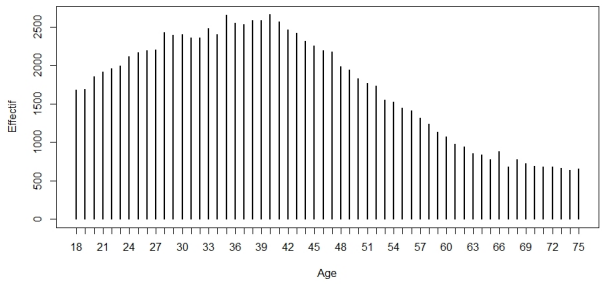

FIGURE 3.1 – *Effectif selon l'âge*

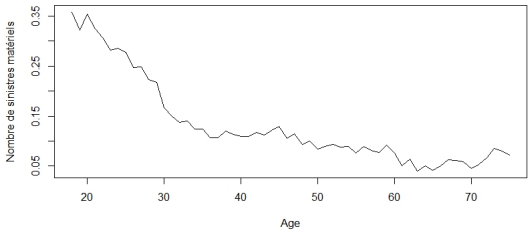

FIGURE 3.2 – *Nombre de sinistres matériels selon l'âge*

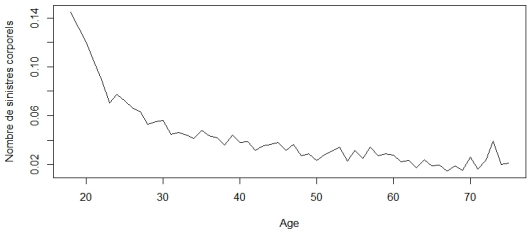

FIGURE 3.3 – *Nombre de sinistres corporels selon l'âge*

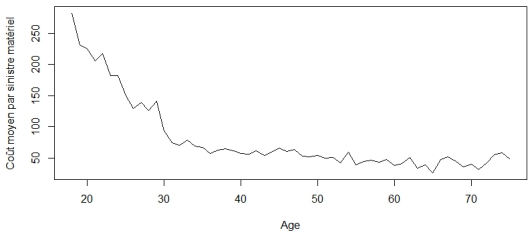

FIGURE 3.4 – *Coût moyen par sinistre matériel selon l'âge*

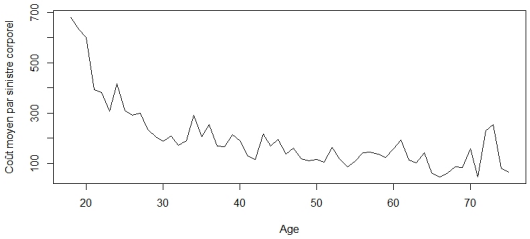

FIGURE 3.5 – *Coût moyen par sinistre corporel selon l'âge*

Le sexe

Il s'agit d'une variable catégorielle qui prend deux valeurs : `Male` (homme) ou `Female` (femme). Dans notre portefeuille, on dénombre 63 443 hommes et 36 578 femmes : il y a donc beaucoup plus d'hommes que de femmes. La sinistralité des femmes est plus faible que celle des hommes, qu'il s'agisse du nombre ou du coût.

FIGURE 3.6 – *Nombre de sinistres matériels selon le sexe*

FIGURE 3.7 – *Nombre de sinistres corporels selon le sexe*

FIGURE 3.8 – *Coût moyen par sinistre matériel selon le sexe*

FIGURE 3.9 – *Coût moyen par sinistre corporel selon le sexe*

La profession

La profession renseigne l'assureur sur l'utilisation du véhicule. Les retraités se distinguent des autres assurés puisqu'ils sont très peu exposés au risque et la charge financière qu'ils représentent est faible. En revanche, les chômeurs constituent la catégorie la plus risquée et la plus coûteuse.

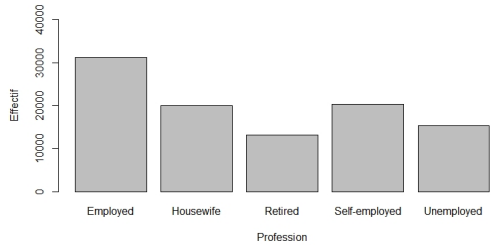

FIGURE 3.10 – *Effectif selon la profession*

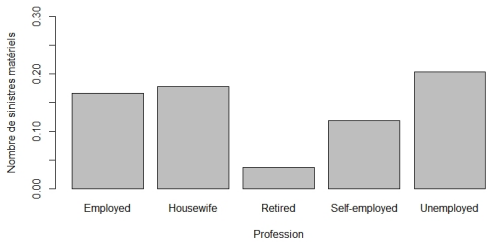

FIGURE 3.11 – *Nombre de sinistres matériels selon la profession*

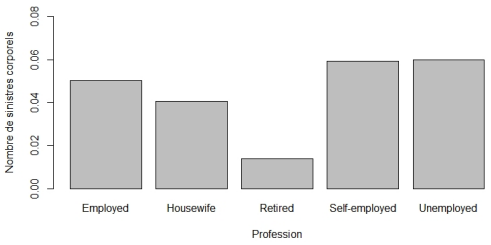

FIGURE 3.12 – *Nombre de sinistres corporels selon la profession*

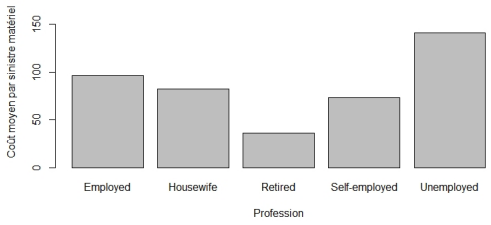

FIGURE 3.13 – *Coût moyen par sinistre matériel selon la profession*

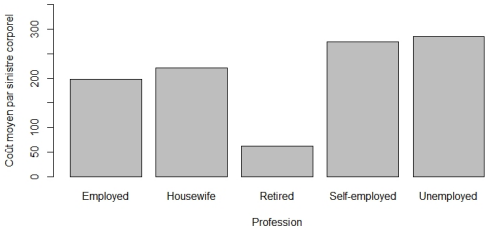

FIGURE 3.14 – *Coût moyen par sinistre corporel selon la profession*

Le bonus

Le bonus du conducteur décrit le profil de risque de l'assuré. Dans notre cas, plus la valeur est élevée, plus l'assuré est risqué. Il est fortement (positivement) corrélé avec le nombre et le coût des sinistres matériels mais a un impact plus limité sur le nombre et le coût des sinistres corporels.

Le portefeuille contient peu de mauvais risques, puisque le nombre d'assurés ayant un malus (valeur > 0) est bien inférieur au nombre d'assurés ayant un bonus (< 0). Les sinistres des mauvais risques sont payés par les bons risques : c'est le principe de la mutualisation sur lequel repose toute activité d'assurance.

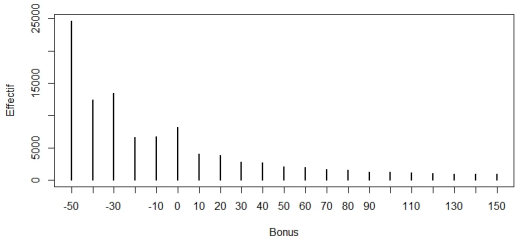

FIGURE 3.15 – *Effectif selon le bonus*

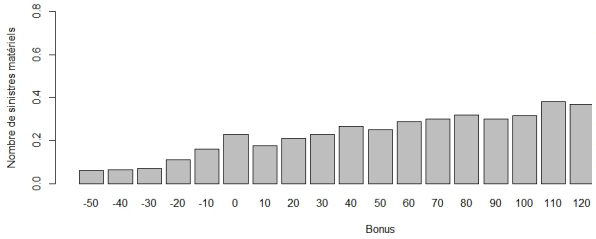

FIGURE 3.16 – *Nombre de sinistres matériels selon le bonus*

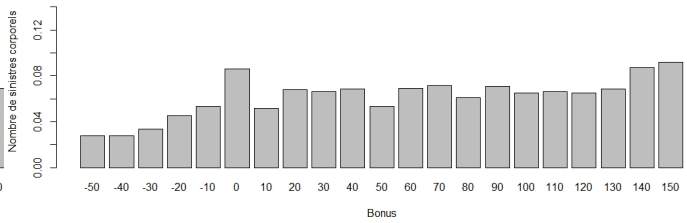

FIGURE 3.17 – *Nombre de sinistres corporels selon le bonus*

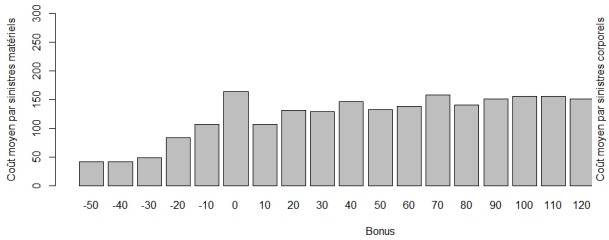

FIGURE 3.18 – *Coût moyen par sinistre matériel selon le bonus*

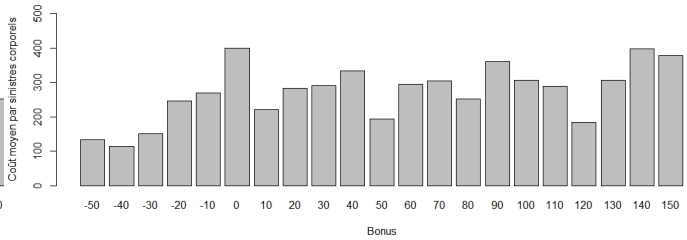

FIGURE 3.19 – *Coût moyen par sinistre corporel selon le bonus*

L'ancienneté d'assurance

L'ancienneté d'assurance n'est pas une variable pertinente dans la mesure où l'historique de la sinistralité des assurés n'est pas disponible.

Elle renseigne néanmoins l'assureur sur la sur-sinistralité du client. Un assuré qui déclare un nombre de sinistres trop important peut faire l'objet d'une résiliation à l'initiative de l'assureur.

Plus l'ancienneté d'assurance de l'assuré est importante, plus le nombre de sinistre observés est faible (rappelons que les sinistres sont observés uniquement sur une année), bien que cela ne soit pas flagrant. On remarque cependant que la variable n'influence pas le montant des sinistres. Les nouveaux contrats représentent 15% du portefeuille et les contrats les plus anciens datent de 15 ans.

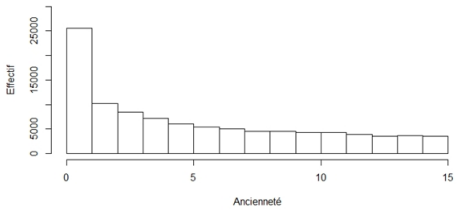

FIGURE 3.20 – *Effectif selon l'ancienneté*

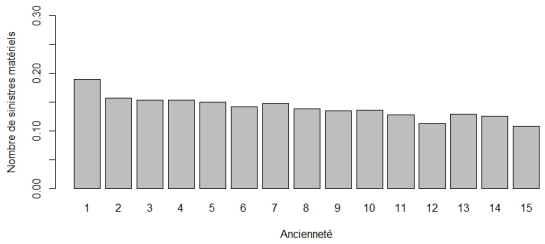

FIGURE 3.21 – *Nombre de sinistres matériels selon l'ancienneté*

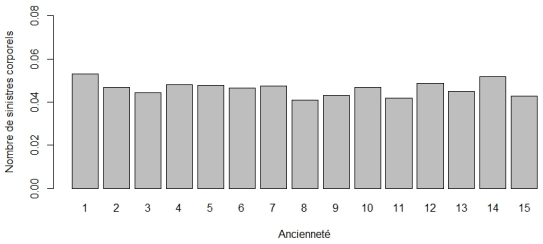

FIGURE 3.22 – *Nombre de sinistres corporels selon l'ancienneté*

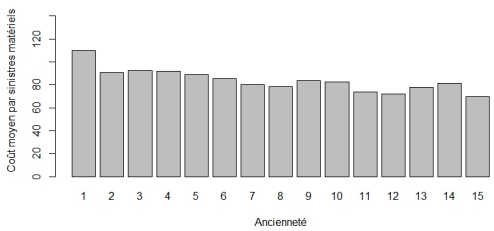

FIGURE 3.23 – *Coût moyen par sinistre matériel selon l'ancienneté*

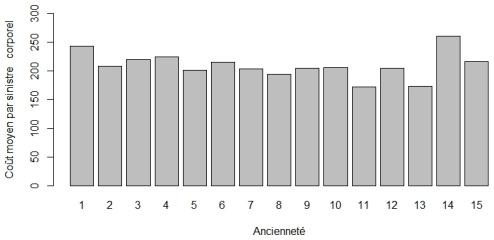

FIGURE 3.24 – *Coût moyen par sinistre corporel selon l'ancienneté*

La région d'habitation, la sous-région d'habitation et la densité

C'est l'information de densité qui intéresse l'assureur à travers ces trois variables.

La variable densité n'est pas expliquée. Il s'agit probablement de la densité de population du lieu de résidence de l'assuré. On peut supposer que plus sa valeur est élevée, plus la densité l'est.

Aucune information n'est donnée concernant les codes attribués aux différentes régions, si ce n'est que l'appartenance à une classe se fait en fonction du profil de risque de la zone géographique.

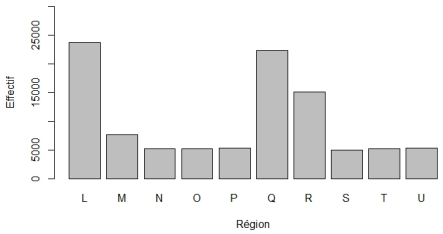

FIGURE 3.25 – *Effectif selon la région*

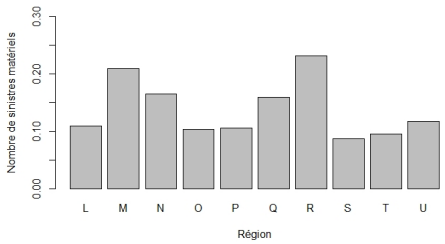

FIGURE 3.26 – *Nombre de sinistres matériels selon la région*

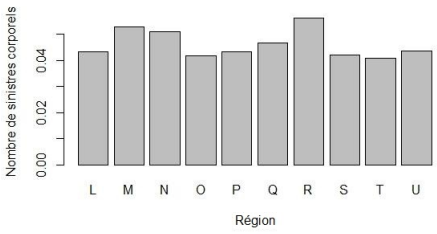

FIGURE 3.27 – *Nombre de sinistres corporels selon la région*

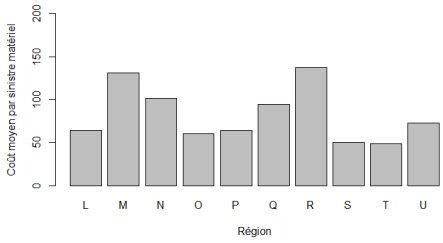

FIGURE 3.28 – *Coût moyen par sinistre matériel selon la région*

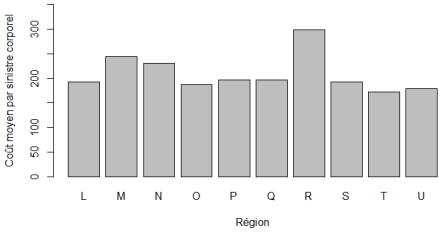

FIGURE 3.29 – *Coût moyen par sinistre corporel selon la région*

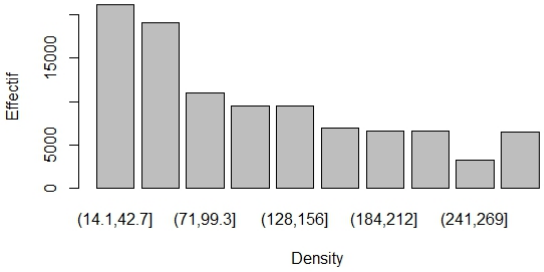

FIGURE 3.30 – *Effectif selon la densité*

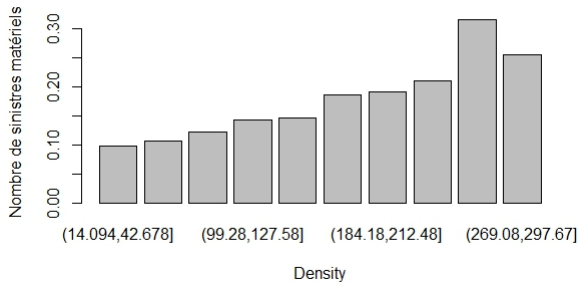

FIGURE 3.31 – *Nombre de sinistres matériels selon la densité*

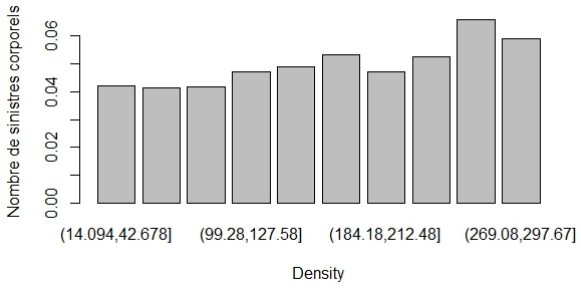

FIGURE 3.32 – *Nombre de sinistres corporels selon la densité*

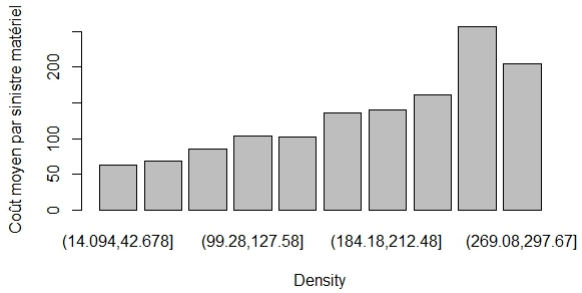

FIGURE 3.33 – *Coût moyen par sinistre matériel selon la densité*

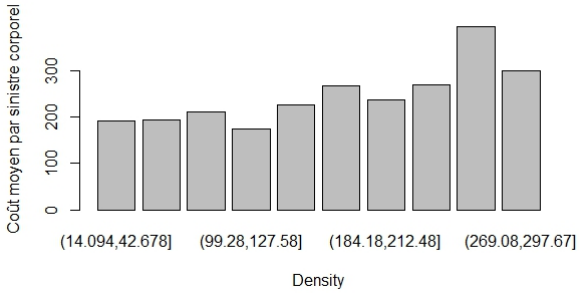

FIGURE 3.34 – *Coût moyen par sinistre corporel selon la densité*

L'exposition

Il s'agit du nombre de jours pendant lesquels le client a été assuré.

Cette variable n'a pas d'impact direct sur le nombre et le coût des sinistres. Elle permet de calculer la fréquence de sinistres, et donc de « normaliser » le nombre de sinistres par assuré. En effet, un assuré qui déclare 2 sinistres en 6 mois n'a pas le même profil de risque qu'un assuré qui déclare 2 sinistres en 12 mois.

73% des assurés ont une durée d'exposition de 365 jours.

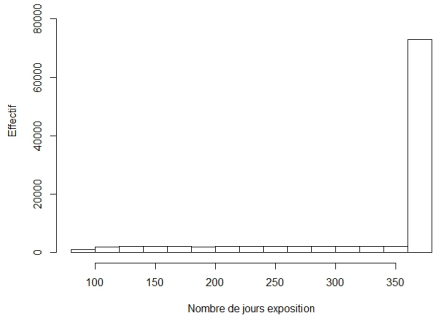

FIGURE 3.35 – *Effectif selon les jours d'exposition*

3.1.2 Variables liées au véhicule

Le type, la catégorie et le groupe

Les caractéristiques du véhicule sont indispensables pour l'assureur.

Les véhicules du portefeuille sont classés par type et groupe sans aucune indication quant aux choix de classification propre à chaque assureur. De la même manière que pour les variables géographiques, l'appartenance à une classe dépend des risques liés aux caractéristiques du véhicule.

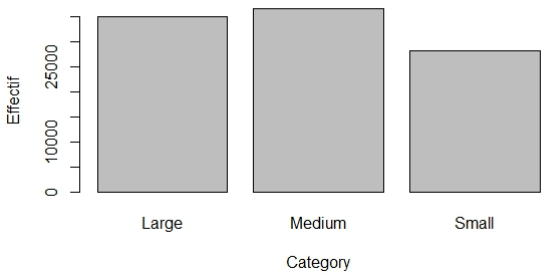

FIGURE 3.36 – *Effectif selon la catégorie*

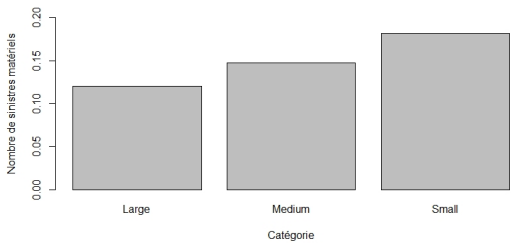

FIGURE 3.37 – *Nombre de sinistres matériels selon la catégorie*

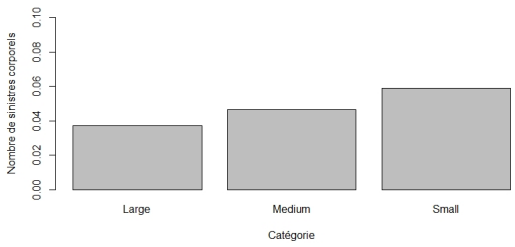

FIGURE 3.38 – *Nombre de sinistres corporels selon la catégorie*

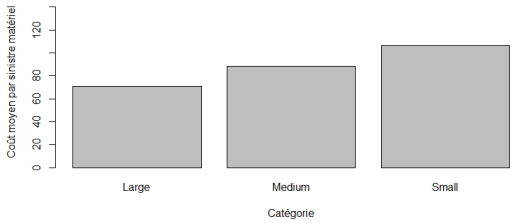

FIGURE 3.39 – *Coût moyen par sinistre matériel selon la catégorie*

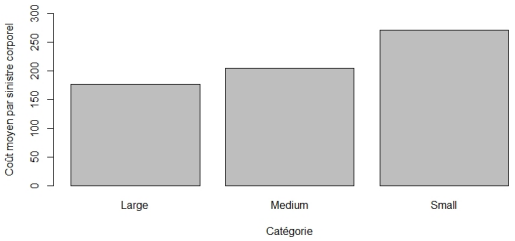

FIGURE 3.40 – *Coût moyen par sinistre corporel selon la catégorie*

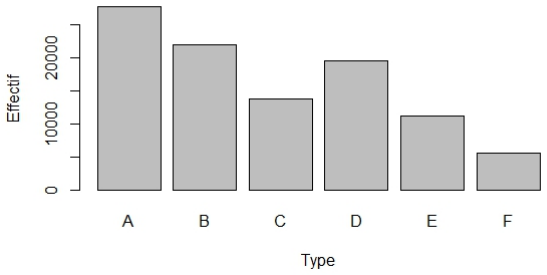

FIGURE 3.41 – *Effectif selon le type*

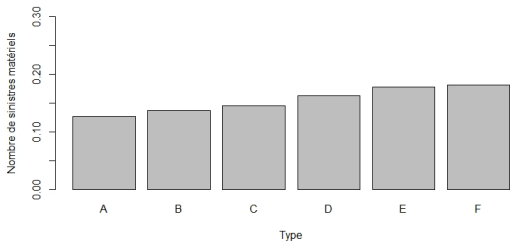

FIGURE 3.42 – *Nombre de sinistres matériels selon le type*

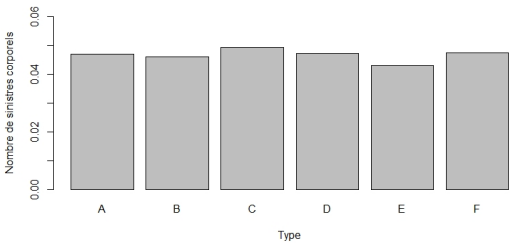

FIGURE 3.43 – *Nombre de sinistres corporels selon le type*

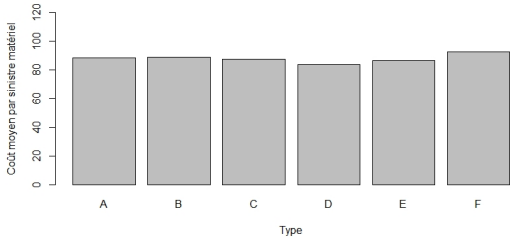

FIGURE 3.44 – *Coût moyen par sinistre matériel selon le type*

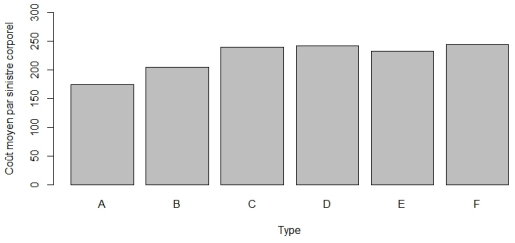

FIGURE 3.45 – *Coût moyen par sinistre corporel selon le type*

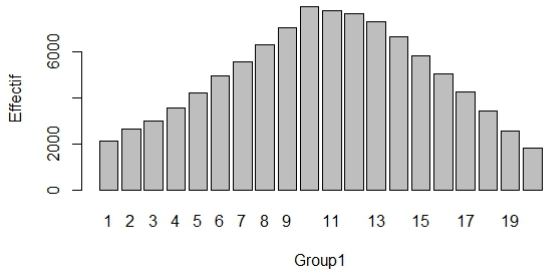

FIGURE 3.46 – *Effectif selon les groupes*

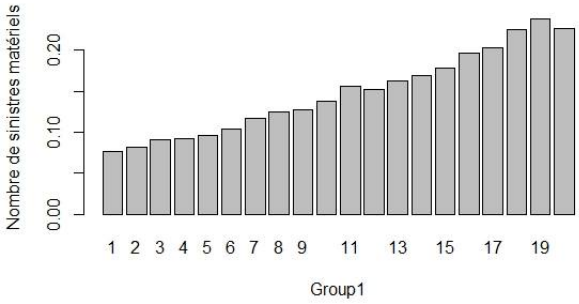

FIGURE 3.47 – *Nombre de sinistres matériels selon les groupes*

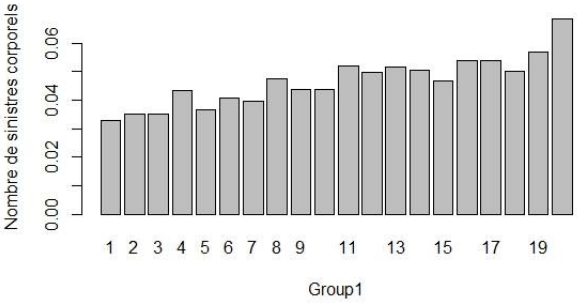

FIGURE 3.48 – *Nombre de sinistres corporels selon les groupes*

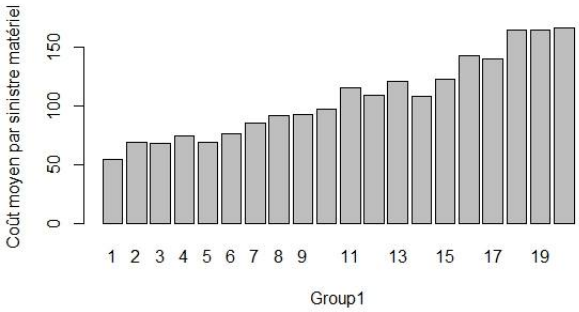

FIGURE 3.49 – *Coût moyen par sinistre matériel selon les groupes*

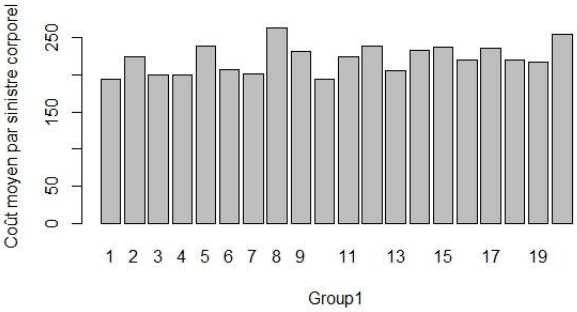

FIGURE 3.50 – *Coût moyen par sinistre corporel selon les groupes*

La valeur

La valeur du véhicule complète les variables type, catégorie et groupe. Les véhicules d'une valeur supérieure à 30 000 € (9% des véhicules) causent le plus grand nombre de sinistres et les sinistres les plus coûteux. Rappelons que la Responsabilité Civile prend uniquement en charge l'indemnisation des tiers, et non celle des assurés.

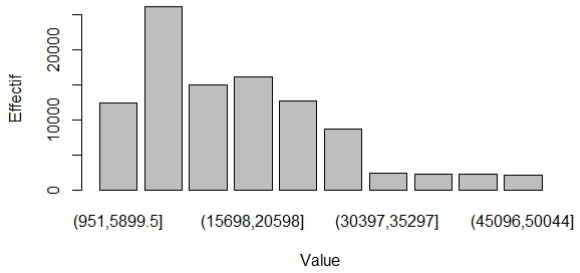

FIGURE 3.51 – *Effectif selon la valeur*

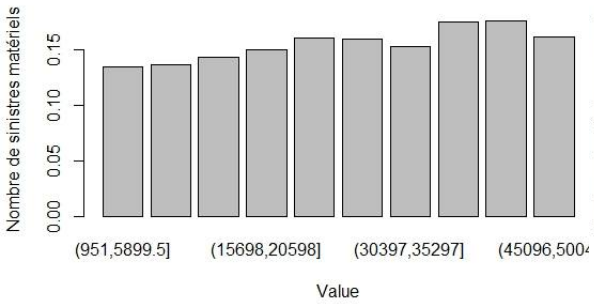

FIGURE 3.52 – *Nombre de sinistres matériels selon la valeur*

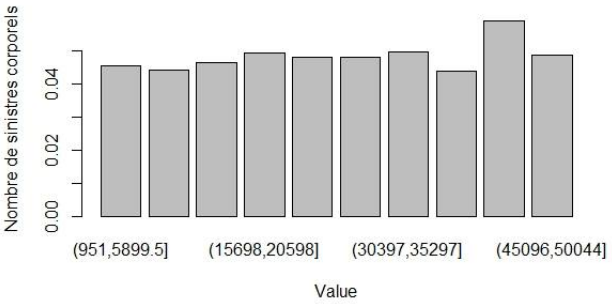

FIGURE 3.53 – *Nombre de sinistres corporels selon la valeur*

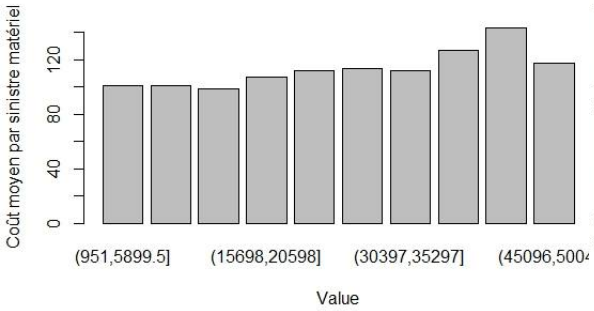

FIGURE 3.54 – *Coût moyen par sinistre matériel selon la valeur*

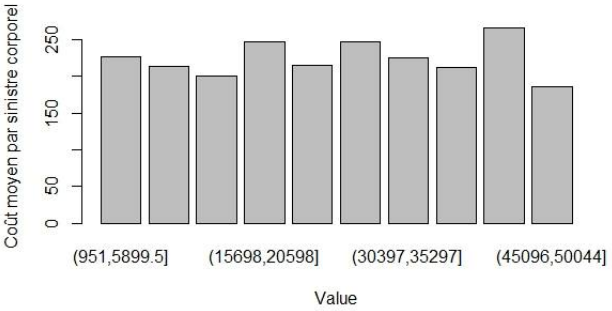

FIGURE 3.55 – *Coût moyen par sinistre corporel selon la valeur*

3.1.3 Variables liées à la souscription de la garantie dommages

Cette variable binaire indique si l'assuré a souscrit ou non à la garantie dommages. Elle assure le véhicule du client. D'après les graphiques 3.56, 3.57, 3.58 et 3.59, les assurés qui y ont souscrit (51%) sont moins exposés au risque que les autres clients.

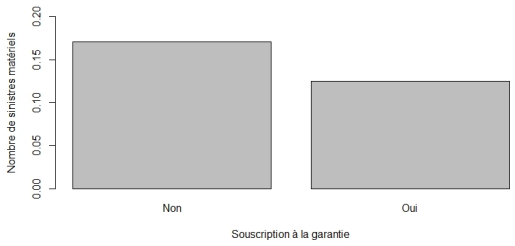

FIGURE 3.56 – *Nombre de sinistres matériels selon la souscription à la garantie*

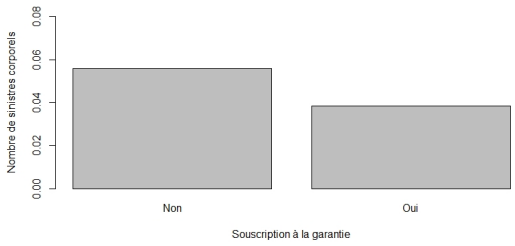

FIGURE 3.57 – *Nombre de sinistres corporels selon la souscription à la garantie*

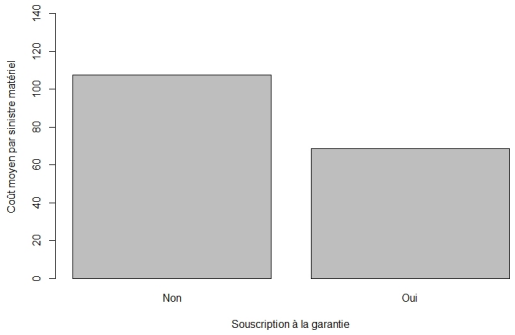

FIGURE 3.58 – *Coût moyen par sinistre matériel selon la souscription à la garantie*

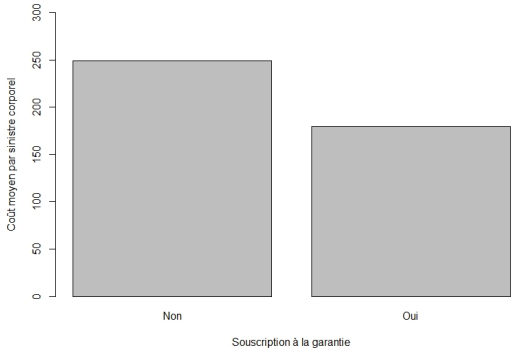

FIGURE 3.59 – *Coût moyen par sinistre corporel selon la souscription à la garantie*

3.1.4 Variables observées a posteriori

Le nombre de sinistres matériels

Le nombre de sinistres matériels d'un assuré est calculé sur la durée d'exposition de celui-ci. Il y a 14 748 sinistres au total, et un nombre moyen de sinistres par assuré de 0,147449. La plupart des assurés n'ont aucun sinistre, et le nombre de sinistres maximum est de 7.

Quantiles	97%	98%	99%	100%
Nombres de sinistres matériels	0	1	2	7

TABLE 3.1 – *Quantiles du nombre de sinistres matériels*

Le nombre de sinistres corporels

Le nombre de sinistres corporels d'un assuré est aussi calculé sur la durée d'exposition de celui-ci. Il est plus faible que le nombre de sinistres matériels : 4 680 sinistres au total, et une moyenne de 0,04679 par assuré. La plupart des assurés n'ont aucun sinistre, et le nombre maximum de sinistres est de 3 :

Quantiles	95%	96%	97%	98%	99%	100%
Nombres de sinistres corporels	0	1	1	1	1	3

TABLE 3.2 – *Quantiles du nombre de sinistres corporels*

Le montant des sinistres matériels

Le montant total des sinistres matériels s'élève à 10 615 730 €, soit une moyenne de 106,14 € par assuré. La plupart des montants de sinistres sont relativement faibles. Parmi les personnes ayant déclaré un sinistre matériel, seuls 1 000 assurés ont coûté plus de 4 661,98 € à l'assureur.

Quantiles	40%	50%	99%	100%
Montant des sinistres matériels	0 €	561,45 €	4 661,98 €	12 878,37 €

TABLE 3.3 – *Quantiles du montant des sinistres matériels*

Le montant des sinistres corporels

Le montant total des sinistres corporels s'élève à 22 280 961 €, soit une moyenne de 222,76 € par assuré. Les sinistres corporels coûtent plus cher à l'assureur que les sinistres matériels : 2 fois plus en moyenne. Parmi les assurés ayant déclaré un sinistre corporel, au moins 10 000 assurés ont coûté plus de 13 000 € à l'assureur.

Quantiles	0%	10%	20%	50%	90%	100%
Montant des sinistres corporels	0 €	78,94 €	322,31 €	2 147,49 €	13 357,63 €	69 068,03 €

TABLE 3.4 – *Quantiles du montant des sinistres matériels*

3.2 Recherche de valeurs aberrantes

Les données contiennent régulièrement des valeurs aberrantes dues à des erreurs de saisie : des conducteurs ayant moins de 18 ans, des bonus inférieurs à 50% ou des anciennetés de contrats supérieures à 80 ans pour des assurés trentenaires.

La vérification est possible pour toutes les variables du portefeuille, excepté pour la densité : l'échelle n'étant pas fournie, il n'est pas possible de savoir à quoi elle correspond.

Remarquons dans un premier temps que les bornes des variables sont cohérentes :

- l'âge varie de 18 à 75 ans
- l'ancienneté de l'assuré varie de 0 à 15 ans
- la valeur du véhicule varie de 1 000 à 49 995 euros
- le nombre de jours d'exposition varie de 91 à 365 jours
- le nombre de sinistres varie de 0 à 7 sinistres
- le montant des sinistres varie de 0 à 69 068,03 euros

Cependant, si on croise entre elles les variables liées à l'âge et à l'ancienneté d'assurance de l'assuré chez l'assureur, le résultat n'est pas cohérent : 12 420 assurés ont moins de 33 ans et ont une ancienneté de 15 ans. Cela signifie qu'ils ont souscrit leur contrat avant l'année de leurs 18 ans, ce qui n'est pas possible. La décision de supprimer ces données a été prise afin de ne pas fausser les résultats avec des données mal reportées.

3.3 Recherche de variables redondantes

La base de données `Training` contient beaucoup de variables. L'analyse de celles-ci peut permettre d'éliminer les variables qui donneraient une information redondante et surchargeraient inutilement le modèle.

La corrélation des variables deux à deux est un bon indicateur. Nous pouvons en effet considérer que deux variables relatant la même information sont redondantes si elles sont très fortement corrélées.

Dans la Figure 3.60, on remarque que la corrélation entre la variable `Group2` et la variable `SubGroup2` concernant toutes deux la région d'habitation est extrêmement élevée comparée aux autres corrélations : 98%.

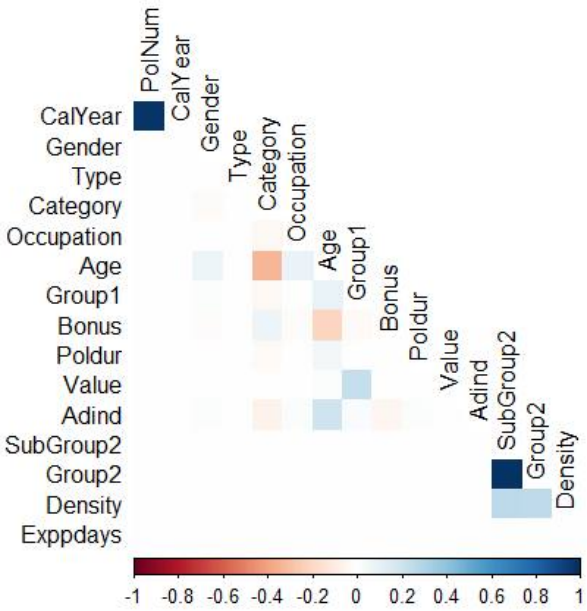

FIGURE 3.60 – *Tableau des corrélations*

L'écartement de la variable `Group2` est donc envisagé afin de ne pas surcharger le modèle.

3.4 Vérification du format des variables

La base de données `Training` regroupe à la fois des variables qualitatives (exprimant une qualité) et des variables quantitatives (exprimant une quantité).

Le logiciel utilisé doit distinguer les deux catégories de variables, au risque de faire des erreurs dans le calibrage des modèles.

Dans notre base de données, les variable `Adind` et `Group1` sont reconnues comme des variables quantitatives, alors qu'il s'agit de variables qualitatives. Il faut donc procéder à une transformation de ces variables à l'aide du logiciel *R*.

Ce traitement préliminaire des données a permis de s'approprier le portefeuille de la base `Training`. 12 420 valeurs jugées aberrantes ont été éliminées, amenant la base de données à 87 601. Une seule variable considérée comme redondante a été écartée. Enfin, deux variables quantitatives ont été transformées en variables qualitatives. La première étape du projet peut alors être initiée.

Chapitre 4

Tarification du produit d'assurance automobile

Rappelons l'intitulé de notre BE : tarification d'un produit d'assurance IARD et analyses d'impacts sur un marché concurrentiel. L'objectif de l'étape de prédiction est la tarification du produit d'assurance automobile de chacun des assurés potentiels de la base test appelée `Pricing`. Cette tarification est représentée par la prime pure pour l'année 2011, estimation de ce qu'un assuré coûtera à l'assureur au cours cette année 2011.

Cette prédiction pourrait être réalisée à partir de l'historique de la sinistralité de ces clients potentiels. Or, cet historique n'est pas présent dans la base `Pricing`. La question se pose donc de savoir comment estimer le nombre de sinistres et le coût moyen d'un sinistre, données indispensables à l'estimation de la prime pure demandée à chaque assuré pour l'année 2011.

La démarche employée dans ce projet consiste à se baser sur les profils des clients actuels du portefeuille de l'assureur disponibles dans la base `Training`, afin de prédire le coût des futurs assurés de la base `Pricing`. Il est en effet raisonnable de penser que deux assurés ayant les mêmes caractéristiques, et donc le même profil de risque, ont une forte probabilité de déclarer le même nombre de sinistres. Ce raisonnement est le pilier central de nos travaux.

La prédiction de la prime pure étant obtenue en multipliant deux variables, nombre de sinistres et coût moyen d'un sinistre, il s'agit, dans la base `Training`, d'établir une relation entre le nombre de sinistres et le coût moyen d'un sinistre (les variables à expliquer) et les caractéristiques de l'assuré dont dispose cette base d'apprentissage (les variables explicatives).

Le logiciel *R*, après qu'on lui ait précisé la méthode retenue, détermine les relations entre les variables explicatives et les variables à expliquer. C'est l'étape de calibrage. Un modèle est ainsi défini.

Le modèle peut alors être appliqué à la base test pour y prédire les variables à expliquer à partir des variables explicatives de cette même base.

La Figure 4.1 résume la démarche de prédiction.

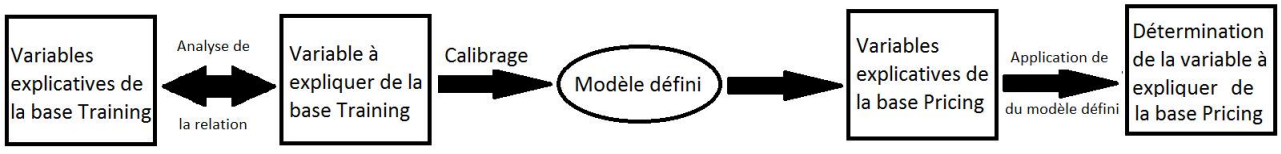

FIGURE 4.1 – *Schéma de la démarche de prédiction*

La prédiction de la tarification du produit d'assurance automobile sera alors réalisée pour les 36 311 assurés potentiels de la base `Pricing`.

4.1 Création des variables à expliquer

Comme souligné précédemment, la relation est la suivante pour calculer la prime pure d'un assuré :

$$Prime\ pure = Nombre\ de\ sinistres \times Coût\ moyen\ par\ sinistre\ (causé\ par\ l'assuré)$$

Le nombre de sinistres par assuré est disponible dans la base `Training`. Cette variable n'est pas exploitable en l'état puisque tous les assurés n'ont pas été assurés sur la même durée.

Aussi est-il nécessaire de normaliser l'occurrence des sinistres. Cette nouvelle variable est obtenue en divisant le nombre de sinistres de l'assuré par son exposition (nombre de jours de couverture de l'assuré), puis en multipliant ce résultat par le nombre de jours dans une année, soit 365.

De la même manière, le montant des sinistres par assuré ne peut pas être utilisé comme une variable à expliquer. La normalisation de cette variable se fait en la divisant par le nombre de sinistres déclarés de l'assuré. Cette nouvelle variable, appelée Coût moyen par sinistre, correspond à la moyenne du montant d'un sinistre pour chaque assuré.

Par ailleurs, une distinction entre les sinistres matériels et les sinistres corporels est essentielle. La probabilité de survenance des sinistres corporels est effectivement plus faible que celle des sinistres matériels, il ne serait donc pas correct de les prédire avec des modèles similaires.

Les quatre variables à expliquer sont donc les suivantes :

- Nombre des sinistres matériels
- Nombre des sinistres corporels
- Coût moyen d'un sinistre matériel
- Coût moyen d'un sinistre corporel

Il est tout aussi nécessaire de distinguer les sinistres attritionnels (classiques) des sinistres graves (atypiques). Le coût de ces derniers étant plus important que celui des sinistres attritionnels, il semble incorrect de les estimer avec les mêmes modèles que les sinistres attritionnels.

Les deux variables à expliquer concernant le coût moyen d'un sinistre nécessitent donc un travail supplémentaire.

En calculant les quantiles de ces deux variables, nous avons observé que 5% des sinistres matériels expliquent 70% du montant total des sinistres matériels et 1% des sinistres corporels expliquent 70% du montant total des sinistres corporels.

L'échantillon présente donc des valeurs extrêmes, c'est-à-dire de sinistres graves. Le calcul du kurtosis permet de vérifier cette intuition.

Le kurtosis mesure l'aplatissement de la distribution de la loi des coûts des sinistres grâce au moment centré d'ordre 4 des coûts des sinistres.

$$\mu_4 = \mathbb{E}\left[(X-m)^4\right]$$

avec X la variable « coût moyen d'un sinistre » et m sa moyenne.

La division de ce terme par la variance au carré permet d'obtenir un coefficient sans unité :

$$kurtosis = \frac{\mu_4}{\sigma^4}$$

Le kurtosis normalisé exprime l'excès d'aplatissement de la variable. On l'obtient en retranchant le nombre 3 au kurtosis. Ce nombre correspond au kurtosis de la loi normale. Plus le coefficient est élevé, plus la queue de la loi de la variable est lourde. C'est-à-dire que les quelques gros sinistres expliquent une grande partie du montant total des sinistres.

Dans notre cas, le kurtosis vaut 93 pour les sinistres matériels et 282 pour les sinistres corporels. La présence de sinistres graves est bien confirmée.

La théorie des valeurs extrêmes permet de déterminer le seuil du montant des sinistres à partir duquel on peut parler de sinistres graves.

Il s'agit d'une théorie complexe qui nécessiterait une explication conséquente. Celle-ci dépasse le cadre de l'étude, nous nous contenterons donc d'une synthèse et de l'interprétation des résultats.

Si néanmoins, le lecteur souhaite approfondir le sujet, il peut se référer au mémoire de William GEHIN, ancien élève de l'EURIA, intitulé *Modélisation des queues de distribution des rendements des actifs financiers. Application à la mesure du risque de marché et à la détermination de stratégies d'investissement*[1].

La théorie des valeurs extrêmes se divise en deux groupes :

- La méthode des maxima par blocs
- La méthode du dépassement de seuil u

Seule la seconde méthode est abordée ici. Elle consiste à définir un seuil et observer le comportement des valeurs qui dépasse ce seuil u.

Soit X l'échantillon observé, le théorème de Fisher-Tippett indique que :

soit $F(x)$ la fonction de répartition de l'échantillon observé s'il existe deux suites réelles $a_n > 0$ et b_n telles que pour n assez grand :

$$F(a_n x + b_n)^n \to H(x)$$

avec $H(x)$ une fonction de répartition non dégénérée, alors $H(x)$ fait nécessairement partie des fonctions de répartitions de la famille :

$Grumbel$:	$H(x) = exp(-exp(-x))$	
$Fréchet$:	$H(x) = exp(-x^{-a})$	$pour\ x > 0,\ 0\ sinon$
$Weibull$:	$H(x) = exp(-\lvert x\rvert^a)$	$pour\ x < 0,\ 1\ sinon$

Ces trois familles de fonction de répartition peuvent être réunit en une seule fonction : la fonction GEV (Generalized Extreme Value) :

$$GEV(\mu, \xi, \phi) = \begin{cases} exp(-(1 + \xi\frac{x-\mu}{\phi})_+^{\frac{1}{\xi}}) & pour\ \xi \neq 0 \\ exp(-exp(\frac{x-\mu}{\phi})) & pour\ \xi = 0 \end{cases}$$

Les valeurs dépassant le seuil sont modélisées par cette loi. On a alors :

$$F^n(x) \approx GEV(\mu, \xi, \phi)$$

Pour la suite, $Y = X - u$ et $F_u(y)$ la fonction de répartition de Y conditionnelle à $Y > 0$:

$$\begin{aligned} F_u(x) &= P[Y \leq y | Y > 0] \\ &= \frac{P[0 < Y \leq y]}{1 - P[Y \leq 0]} \\ &= \frac{P[u < X \leq u + y]}{1 - P[X \leq u]} \\ &= \frac{F(u+y) - F(u)}{1 - F(u)} \end{aligned}$$

Or $F^n \approx GEV(\mu, \xi, \phi)$ ce qui équivaut à :

$$log(F(x)) = \frac{-1}{n}(1 + \xi\frac{x-\mu}{\phi})^{-\frac{1}{\xi}}$$

L'analyse est intéressante pour n grand, $log(F^n(x))$ peut donc être approximé par $-(1 - F(x))$:

$$1 - F(x) \approx \frac{1}{n}(-(1 + \xi\frac{x-\mu}{\phi})^{-\frac{1}{\xi}}$$

1. ce mémoire a servi de support pour réaliser la distinction entre les deux types de sinistres.

$F_u(y)$ équivaut à :

$$\frac{F(u+y)-F(u)}{1-F(u)}$$
$$= 1 - \frac{1-F(u+y)}{1-F(u)}$$
$$\approx 1 - (\frac{1+\xi\frac{u+y-\mu}{\phi}}{1+\xi\frac{u-\mu}{\phi}})^{-\frac{1}{\xi}}$$
$$\approx 1 - (1+\frac{\xi}{\beta}y)^{-\frac{1}{\xi}}$$

avec $\beta = \phi + \xi(u-\mu)$

Il s'agit de la fonction de répartition GPD (Generalized Pareto Distribution). Les valeurs situées au-dessus du seuil suivent cette loi.

La fonction de répartition de la loi s'écrit aussi sous cette forme :

$$F_{\xi,\beta} = \begin{cases} 1-(1+\xi\frac{x}{\beta})^{-\frac{1}{\xi}} & pour\ \xi \neq 0 \\ 1-exp(\frac{-x}{\beta}) & pour\ \xi = 0 \end{cases}$$

Elle est définie par deux paramètres :

- ξ le paramètre de forme. C'est le plus important car il définit l'épaisseur de la queue de la loi.
- β le paramètre d'échelle.

Le choix du seuil

C'est un choix relativement subjectif qui s'appuie sur deux graphiques :

- Le Mean Residual Life Plot : il représente le seuil u en fonction de l'excès moyen (moyenne des valeurs au-dessus de $u-u$). Le graphe doit être linéaire en-dessous du seuil.
- Le Parameter Stability Plot : il représente la valeur des paramètres estimés de la loi GPD en fonction de u. Le graphe doit être stable (valeur du paramètre constante) en-dessous du seuil.

Sinistres matériels

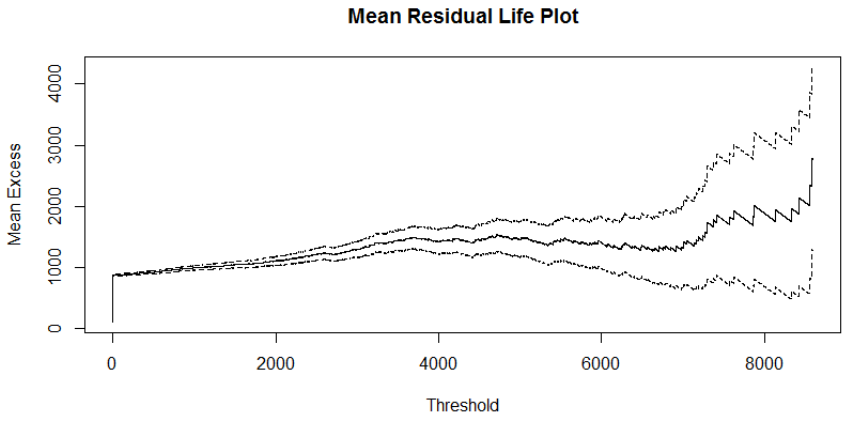

FIGURE 4.2 – *Mean Residual Life Plot (matériel)*

Le graphe 4.2 est linéaire en dessous de la valeur 4 000. Les deux graphes 4.3 sont stables en dessous de cette valeur. Les sinistres matériels dont le coût moyen est supérieur à 4 000 € sont considérés comme graves.

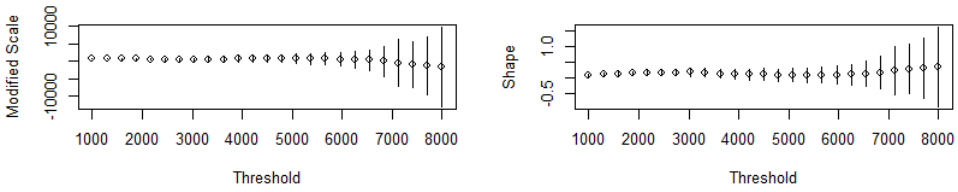

FIGURE 4.3 – *Threshold Choice Plot (matériel)*

Sinistres corporels

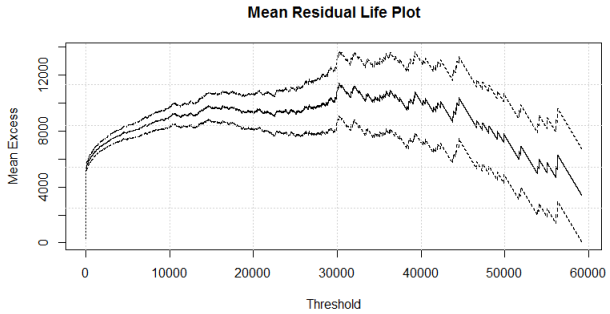

FIGURE 4.4 – *Mean Residual Life Plot (corporel)*

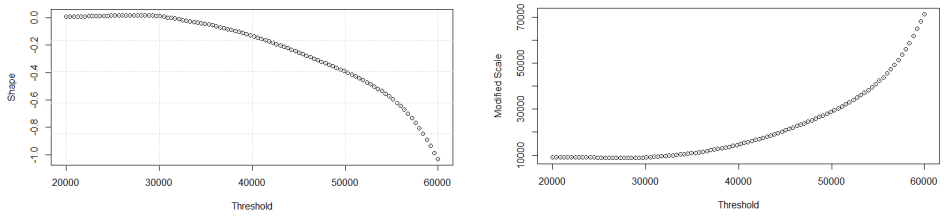

FIGURE 4.5 – *Threshold Choice Plot (corporel)*

De la même manière, on fixe le seuil des sinistres corporels sur les graphes 4.4 et 4.5 à 30 000 €. Les sinistres ayant un coût supérieur sont considérés comme graves.

4.2 Prédiction des variables à expliquer (sinistres attritionnels)

4.2.1 La régression linéaire

Il s'agit du modèle le plus simple. Il repose sur une hypothèse très forte : la relation qui lie les variables explicatives et la variable à expliquer est linéaire. C'est-à-dire que la variable à expliquer des assurés y s'écrit comme une combinaison linéaire des j variables explicatives x_j de ces assurés. L'équation linéaire obtenue est celle d'une droite qui passe au mieux par les points :

$$y = a_0 + \sum_{j=1}^{n} a_j x_j + \epsilon$$

où les a_j sont les coefficients de la j^{eme} valeur à prédire. Ils sont déterminés lors du calibrage du modèle, et ϵ est l'erreur du modèle.

En réécrivant la relation linéaire sous forme matricielle on obtient :

$$Y = aX + \epsilon$$

Trois hypothèses portant sur les erreurs doivent être vérifiées pour appliquer ce modèle à un échantillon :

- Leur variance σ^2 est constante, c'est-à-dire que l'erreur ne dépend pas des variables explicatives
- Elles suivent une loi normale $N(0, \sigma^2)$
- Elles sont indépendantes

On peut en déduire une hypothèse portant sur la variable à expliquer : y suit une loi normale $N(a_0 + \sum_{j=1}^{n} a_j x_j, \sigma^2)$ et sont indépendants.

Une fois les paramètres estimés, le modèle s'écrira :

$$\hat{y} = \hat{a}_0 + \sum_{j=1}^{n} \hat{a}_j x_j + \hat{\epsilon}$$

Les résidus estimés se calculent ainsi : $\hat{\epsilon} = y - \hat{y}$ avec $\hat{y}$ la valeur prédite de y.
Les coefficients sont déterminés afin de minimiser la somme des carrés des résidus.

$$\min \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

La solution est le vecteur a tel que : $a = (X'X)^{-1}X'Y$ avec X' la transposée de X.

Prédiction du nombre de sinistres

Le calibrage du modèle se fait uniquement sur les assurés ayant déclaré au moins 1 sinistre. En effet, l'assureur ne peut jamais être certain qu'un assuré ne provoquera aucun sinistre.

L'implémentation de la méthode de régression linéaire sur R se fait de la manière suivante :

```
##Sinistres matériels

#etape1
fit = lm (nbMat ~.,data=Training) #lm = régression linéaire
                                  #data = base d'apprentissage
#etape2
nbMat = predict(fit, newdata = Pricing) #predict = étape de prédiction
                                        #newdata = base de test

##Sinistres corporels

#etape1
fit = lm (nbCor ~.,data=Training) #lm = régression linéaire
                                  #data = base d'apprentissage
#etape2
nbCor = predict(fit, newdata = Pricing) #predict = étape de prédiction
                                        #newdata = base de test
```

La première étape correspond au calibrage du modèle tandis que la deuxième correspond à l'application du modèle sur les variables explicatives de la base de test.
Les éléments permettant d'évaluer la pertinence des variables explicatives et la performance du modèle sont obtenus grâce à la fonction *summary*.
La pertinence des variables explicatives est vérifiée par la p-value. A chaque variable explicative est en effet associée à cet indicateur.

Dans ce modèle, la p-value de toutes les variables explicatives est inférieure à 5%. Les variables explicatives ont donc toutes un impact significatif sur la variable à expliquer.

La performance du modèle de régression linéaire peut être quant à elle évaluée à l'aide du R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

avec y_i la variable prédite et $\bar{y}$ la moyenne de la valeur réelle.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.2192080	0.1794868	-23.507	< 2e-16	***
Gender2	0.3093459	0.0297026	10.415	< 2e-16	***
Type2	0.0381451	0.0414851	0.919	0.357840	
Type3	0.1590140	0.0460796	3.451	0.000559	***
Type4	0.2273332	0.0404426	5.621	1.90e-08	***
Type5	0.3112078	0.0467784	6.653	2.88e-11	***
Type6	0.3588753	0.0583525	6.150	7.74e-10	***
Category2	-0.0186508	0.0349398	-0.534	0.593482	
Category3	0.0719974	0.0358453	2.009	0.044584	*
Occupation2	0.3228790	0.0367557	8.784	< 2e-16	***
Occupation3	-0.9606275	0.0912627	-10.526	< 2e-16	***
Occupation4	-0.0813289	0.0414464	-1.962	0.049731	*
Occupation5	0.2747105	0.0375986	7.306	2.74e-13	***

FIGURE 4.6 – *p-value des variables explicatives*

Il présente néanmoins un inconvénient : il augmente avec le nombre de variables. Or la robustesse du modèle diminue avec l'augmentation de ce nombre. C'est pour cela qu'il est préférable d'utiliser le R^2 ajusté qui lui prend en compte le nombre de variables du modèle.

$$R^2 \ ajusté = R^2 - \frac{k\left(1-R^2\right)}{n-k-1} = 1 - \frac{\left(1-R^2\right)(n-1)}{n-k-1}$$

avec n le nombre de variables et k le nombre de variables explicatives.

Variables à expliquer	R^2 ajusté
Fréquence des sinistres matériels	0,1111
Fréquence des sinistres corporels	0,01034

TABLE 4.1 – *Tableau des R^2 ajustés du nombre de sinistres*

La régression linéaire n'est à ce stade pas la méthode la plus adaptée à notre cas (R^2 très inférieur à 1). L'hypothèse de linéarité est trop forte pour l'échantillon étudié. Il a néanmoins été décidé de le conserver dans la suite du projet, dans l'optique de comparer suffisamment de modèles.

Prédiction du coût moyen d'un sinistre attritionnel

La prédiction concernant pour le moment les sinistres attritionnels, le calibrage des modèles se fait uniquement sur les sinistres dont le coût est compris entre 0 et 4 000 € pour les sinistres matériels ou 30 000 € pour les sinistres corporels [2].

L'implémentation de la méthode de régression linéaire sur R se fait de la même manière que pour la modélisation du nombre de sinistres :

```
##Sinistres matériels

#etape1
fit = lm (coutMoyMat ~.,data=Training)
#etape2
coutMoyMat = predict(fit, newdata = Pricing)

##Sinistres corporels

#etape1
fit = lm (coutMoyCor ~.,data=Training)
#etape2
coutMoyCor = predict(fit, newdata = Pricing)
```

2. seuil déterminé par la théorie des valeurs extrêmes page 29.

Pour la régression de ces deux variables à expliquer, les p-values de toutes les variables explicatives sont inférieures à 5%. Elles ont donc toutes un impact significatif sur la variable à expliquer.

Variables à expliquer	R^2 ajusté
Coût moyen d'un sinistre matériel	0,06841
Coût moyen d'un sinistre corporel	0,03553

TABLE 4.2 – *Tableau des R^2 ajustés du coût moyen d'un sinistre*

Ici encore, le R^2 est très faible. Il a été décidé de le conserver pour la raison évoqué précédemment.

4.2.2 Les modèles linéaires généralisés (GLM)

Il s'agit d'une version plus souple de la régression linéaire. La variable à expliquer est transformée par une fonction appelée la fonction lien. C'est la variable transformée qui s'écrit comme une combinaison linéaire des variables explicatives.

$$f(y) = a_0 + \sum_{j=1}^{n} a_j x_j + \epsilon$$

avec f la fonction lien.

Distribution des variables

Test de Kolmogorov-Smirnov

La méthode du GLM nécessitant la précision de la loi de l'échantillon, nous devons comparer la distribution des variables à expliquer à des lois connues. Le critère de Kolmogorov-Smirnov permet d'analyser l'adhésion d'une variable à expliquer à une loi usuelle. C'est un test qui s'appuie sur l'écart entre la fonction de répartition empirique de la variable à expliquer et la fonction de répartition de la loi théorique.

$$\Delta_n = \sup_{x \in \mathbb{R}} \mid F_n(x) - F(x) \mid$$

avec $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}$ la fonction de répartition empirique et $F(x)$ la fonction de répartition de la loi testée.

Le nombre des sinistres suit généralement une loi de Poisson ou une loi binomiale négative. Les tests dont dispose le logiciel ne permettent pas de vérifier cette hypothèse dans le cas de lois discrètes. Il sera donc par la suite admis que les échantillons suivent ces deux lois. Ce test est en revanche applicable aux lois continues. Aussi a-t-il été appliqué au coût moyen d'un sinistre.

Le coût moyen d'un sinistre par assuré suit généralement une loi gamma ou une loi log-normal. Les résultats de la Table 4.15 obtenus avec le test de Kolmogorov-Smirnov montrent les p-value obtenues.

	Comparaison à la loi gamma	Comparaison à la loi log-normale
Coût moyen d'un sinistre matériel par assuré	One-sample Kolmogorov-Smirnov test data: training_ks$Indtppd D = 0.0041, p-value = 0.0748 alternative hypothesis: two-sided	One-sample Kolmogorov-Smirnov test data: training_ks$Indtppd D = 0.8166, p-value < 2.2e-16 alternative hypothesis: two-sided
Coût moyen d'un sinistre corporel par assuré	One-sample Kolmogorov-Smirnov test data: training_ks$Indtpbi D = 0.0044, p-value = 0.04368 alternative hypothesis: two-sided	One-sample Kolmogorov-Smirnov test data: training_ks$Indtpbi D = 0.8172, p-value < 2.2e-16 alternative hypothesis: two-sided

TABLE 4.3 – *Comparaison du coût moyen d'un sinistre avec une loi de Poisson et une loi binomiale négative*

La valeur de la p-value étant supérieure à 5% concernant les sinistres matériels, et légèrement inférieure à 5% pour les sinistres corporels, on retiendra l'hypothèse que ces deux échantillons suivent une loi gamma.

A contrario, cette hypothèse est infirmée en ce qui concerne la loi log-normal, qu'il s'agisse des coûts des sinistres matériels ou des sinistres corporels.

Résultat du test

Les calculs de la moyenne et de la variance empirique permettent de déterminer les paramètres des différentes lois.

Ceux de la loi gamma sont égaux à :

$$shape = \frac{E^2[X]}{Var(X)} \qquad et \qquad scale = \frac{E[X]}{Var(X)}$$

Variables à expliquer	Loi et paramètres
Coût moyen d'un sinistre matériel	Gamma de paramètres $shape = 0,05689793$ et $scale = 0,000064$
Coût moyen d'un sinistre corporel	Gamma de paramètres $shape = 0,0143525$ et $scale = 0,00053$

TABLE 4.4 – *Lois des variables à expliquer*

Implémentation du modèle

Prédiction du nombre de sinistres

L'implémentation du modèle sur le logiciel R se fait de la manière suivante :

```
##Sinistres matériels

#etape1
fit = glm (nbMat ~.,data=Training,family = 'Poisson') #glm = modèle linéaire généralisé
                                                      #family=loi de la distribution
#etape2
nbMat = predict(fit, newdata = Pricing)

##Sinistres corporels

#etape1
fit = glm (nbCor ~.,data=Training,family = 'Poisson')
#etape2
nbCor = predict(fit, newdata = Pricing)
```

Si la variable à expliquer suit une loi de Poisson ou une loi binomiale négative, la fonction lien est la fonction logarithme.

Tout comme pour le modèle de régression linéaire, la p-value de toutes les variables explicatives sont inférieures à 5%, elles ont donc toutes un impact significatif sur la fréquence des sinistres.

L'indice de performance du modèle GLM est l'AIC (Critère d'information d'Akaike). Il se calcule de la manière suivante :

$$AIC = 2k - 2ln(L)$$

avec k le nombre de paramètres du modèle et L la fonction du maximum de vraisemblance :

$$L(a, \sigma|y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2\sigma^2}(y - Xa)^t(y - Xa)\right)$$

avec y le vecteur contenant les y_i, σ^2 leur variance et a le vecteur des coefficients de l'équation linéaire.

Plus l'AIC est faible, meilleur est le modèle. L'AIC n'a pas de borne supérieure.

La Table 4.5 permet de conclure que le modèle le plus performant est le modèle binomial négatif.

	AIC GLM Poisson	AIC GLM binomial négatif
Fréquence des sinistres matériels	83 456	27 335
Fréquence des sinistres corporels	92 332	13 360

TABLE 4.5 – *AIC du nombre de sinistres avec une loi de Poisson et binomiale négative*

Prédiction du coût des sinistres

L'implémentation du modèle est réalisée de la manière suivante :

```
##Sinistres matériels

#etape1
fit = glm (coutMoyMat ~.,data=Training,family = 'Gamma')
#etape2
coutMoyMat = predict(fit, newdata = Pricing)

##Sinistres corporels

#etape1
fit = glm (coutMoyCor ~.,data=Training,family = 'Gamma')
#etape2
coutMoyCor = predict(fit, newdata = Pricing)
```

Si l'échantillon suit une loi gamma, la fonction lien est la fonction inverse.

La p-value de toutes les variables explicatives sont inférieures à 5%, elles ont donc toutes un impact significatif sur la fréquence des sinistres.

	AIC GLM gamma
Coût moyen d'un sinistre matériel	54 275
Coût moyen d'un sinistre corporel	19 937

TABLE 4.6 – *AIC du coût moyen avec une loi de gamma*

En l'absence de comparaison, on ne peut pas tirer de conclusion des valeurs d'AIC obtenues. Le modèle GLM donne de bons résultats, il a été décidé de le conserver pour la suite du projet.

Jusqu'à présent, les modèles implémentés reposent sur l'hypothèse de linéarité de la relation entre la variable à expliquer et les variables explicatives. Celle-ci n'est pas systématiquement vérifiée. L'utilisation des modèles d'apprentissage statistique, ou machine learning, est alors envisageable.

La démarche utilisée dans les prochains modèles est différente : en effet, le machine learning permet uniquement de prédire les variables qualitatives. Dans la suite du projet, la valeur de la variable « Nombre de sinistres » ne correspond plus à un « nombre » mais à une classe : la classe 2 correspond à la classe des assurés ayant 2 sinistres. De même, la continuité de la variable de coût moyen par sinistre ne permet pas d'utiliser les méthodes de prédiction qui suivent. Seul le nombre de sinistre sera simulé par ces modèles. Le coût moyen d'un sinistre sera simulé par un GLM gamma (vu précédemment).

La prédiction du nombre de sinistres en une seule étape donne de mauvais résultats. Il est alors nécessaire de décomposer le travail en $N-1$ étapes, avec N le nombre de vecteurs nombre de sinistres prédits possibles.
A chaque étape, une nouvelle base d'apprentissage est créée, ainsi qu'un nouveau vecteur à prédire. Les valeurs prises par ce vecteur sont 0 ou 1.

On note x la valeur à prédire. Chaque base d'apprentissage créée contient l'ensemble des assurés ayant au moins x sinistres. La valeur 0 est affectée aux assurés ayant déclaré un nombre de sinistres égal à x tandis qu'on associe la valeur 1 aux assurés ayant au moins $x+1$ sinistres. Concrètement, si l'on souhaite prédire les assurés qui ont 3 sinistres, la valeur à prédire est 3. La base d'apprentissage correspondant à cette étape de l'algorithme est composée de tous les assurés ayant au moins 3 sinistres. Le vecteur à prédire prend la valeur 0 pour les assurés ayant déclaré 3 sinistres, et 1 pour ceux ayant déclaré au moins 4 sinistres.

4.2.3 Les réseaux de neurones

Cet algorithme s'inspire du fonctionnement biologique du cerveau. Les variables explicatives sont reliées à des « neurones » par des poids synaptiques. Ceux-ci sont modifiés durant l'apprentissage dans le but d'améliorer la prédiction. Les réseaux de neurones peuvent être utilisés avec deux fonctions de combinaison différentes :

- La fonction linéaire : multiplication des variables explicatives par les différents poids synaptiques associés. La somme pondérée est comparée à une valeur seuil. Si elle est inférieure, le neurone est inactif (affectation de la valeur 0). Si elle est supérieure, le neurone est actif (affectation de la valeur 1).
- La fonction softmax : classification des individus en catégories à la sortie du réseau selon une règle de décision : le seuil. L'affectation d'une valeur se fait de la même manière que celle décrite précédemment.

Dans notre cas, le réseau de neurone utilise la sortie softmax. Il possède 20 neurones dans la couche cachée et le nombre d'itérations est de 200. Nous avons utilisé le seuil par défaut de R.

```
library(nnet) #chargement de la library permettant l'utilisation des réseaux de neurones

type=class.ind(S01)

mod=nnet(type~Age+Density+Bonus
         +Ind.dens1+Ind.dens2+Ind.dens3+Ind.dens4
         +Ind.age1+Ind.age2+Ind.age3+Ind.age4
         +Ind.bonus1+Ind.bonus2+Ind.bonus3+Ind.bonus4,data=z,subset=samp,
         softmax=TRUE,size=20,maxit=200)

nnapp <- predict(mod,z[-samp,],type="class")
```

La performance des modèles d'apprentissage statistique est mesurée à l'aide d'une table comparant la variable à prédire et le vecteur de prédiction. Elle se présente sous la forme de la Table 4.7

	Vecteur de prédiction	
Variable à prédire	VP	FN
	FP	VN

TABLE 4.7 – *Performance des modèles d'apprentissage statisitique*

VP (vrais positifs) représente le nombre d'individus prédits à 0 et qui le sont réellement,
FP (faux positifs) représente le nombre d'individus prédits à 0 mais dont la valeur réelle est 1,
FN (faux négatifs) représente le nombre d'individus prédits à 1 mais dont la valeur réelle est 0,
VN (vrais négatifs) représente le nombre d'individus prédits à 1 et qui le sont réellement.

Les algorithmes sélectionnés doivent présenter un taux de faux positifs et de faux négatifs équivalents, afin de ne pas surestimer ni sous-estimer les coûts.

A partir de ces données, il est possible de calculer :

- Le taux de performance : $T_P = \frac{VN}{VN+FN}$. Il est égal au nombre d'individus prédits correctement à 1 (VN), divisé par l'ensemble des individus dont la valeur de prédiction est 1 (VN et FN).
- Le taux d'erreur globale : $T_{eg} = \frac{FN+FP}{n}$, où n est le nombre d'individus de la base d'apprentissage. Il représente le taux d'individus mal classés.

	Vecteur de prédiction	
Variable à prédire	43 653	183
	5 943	231

TABLE 4.8 – *Table de performance pour le modèle de réseaux de neurones*

Le logiciel R renvoie la Table de performance 4.8 pour le modèle de réseaux de neurones.

Le taux de faux négatif est cinquante fois plus faible que le taux de faux positif. Les prédictions de sinistres sont fortement sous-estimées. Cet algorithme ne peut donc pas être utilisé dans cette étude.

4.2.4 Le Support Vector Machine

Le Support Vector Machine ou Séparateur à Vastes Marges (SVM) est une méthode de classification. Elle consiste à déterminer un hyperplan qui sépare au mieux les données, selon la valeur binaire (0 ou 1) de la variable à prédire. La prédiction affecte les vecteurs d'entrées (vecteurs des variables explicatives) à une classe, qui dépend de leur position par rapport à l'hyperplan mis en place.

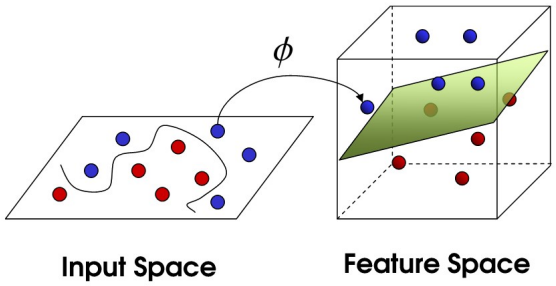

FIGURE 4.7 – *Création de l'hyperplan du modèle SVM*

Lorsque les données ne sont pas linéairement séparables, la première étape de cette méthode consiste à transformer les variables à l'aide d'une fonction noyau, qui ramène les vecteurs d'entrées dans un sous ensemble où ils sont linéairement séparables. La fonction de noyau, notée k s'écrit :

$$k : (x, x') \to \Phi(x).\Phi(x')$$

Où $\Phi : \mathbb{R}^n \to H$ est une fonction qui a pour espace de départ les n variables explicatives et pour espace d'arrivée le sous espace vectoriel H dans lequel les données sont linéairement séparables.

La library $e1071$ de R propose une fonction SVM qui utilise une fonction noyau. Cette fonction noyau est déterminée par l'utilisateur du logiciel et appartient à une sélection de fonctions : fonction linéaire, fonction polynomiale, fonction radiale, fonction sigmoïde ou fonction tangente hyperbolique.

Lors de la modélisation du SVM avec les données de la base `Training`, le meilleur taux de performance s'obtient avec la fonction sigmoïde :

$$(x', y) \to \tanh(\alpha \langle x'|y\rangle + \beta)$$

Où x' (transposée de x) et y représentent deux lignes de la base `Training`, α et β des réels et $\langle .|. \rangle$ est un produit scalaire.

Lors de chaque prédiction, une optimisation de cette fonction est réalisée sur les variables α et β afin d'améliorer les performances de prédiction. L'optimisation de cette fonction permet d'augmenter le taux de performance de 5%.

Dans un souci d'optimisation de performance du modèle, seules les variables les plus corrélées au nombre de sinistres sont conservées : il s'agit de l'âge, du bonus, du groupe du véhicule et de la densité (voir corrélation

page 27).

```
library(e1071)
set.seed(14515)
tablesvm=training[c(8,9,10,12)]

iO=which(training$Numtppd>=1)

SOplus=c(rep(0,dim(training)[1]))
SOplus[iO]=1

samp=echantillon(training$Numtppd)
modelOplus <- svm(tablesvm,SOplus,type="C-classification"
                  , kernel="sigmoid",subset=samp,gamma=0.5,coef0=0.5)
Numtppd=predict(modelOplus,tablesvm[-samp,]) #prédiction
```

	Vecteur de prédiction	
Variable à prédire	38 769	5 021
	5 101	1 120

TABLE 4.9 – *Table de performance du nombre de sinistres matériels pour le modèle SVM*

	Vecteur de prédiction	
Variable à prédire	45 615	2 172
	2 110	114

TABLE 4.10 – *Table de performance du nombre de sinistres corporels pour le modèle SVM*

Le taux de performance de la prédiction du nombre de sinistres matériels est de 18% tandis que celle du nombre de sinistres corporels est plus faible : 5%. Le taux d'erreur global est élevé : 20%. Le modèle est tout de même conservé pour la suite du projet.

4.2.5 Les arbres de décisions

L'arbre y possède une racine (la base d'apprentissage) et un ensemble de nœuds. Le but de cet algorithme est de partitionner la population mère de chaque nœud en plusieurs sous-populations filles. Ce partitionnement est réalisé en fonction des valeurs des variables explicatives. Un critère d'arrêt est fixé. Lorsqu'il est atteint, l'algorithme s'arrête et les populations des derniers nœuds sont affectées à une valeur de la variable à expliquer.

Pour les raisons évoquées précédemment, seules les 4 variables les plus corrélées au nombre de sinistres sont conservées.

```
iO=which(training$Numtppd>=1)

SOplus=c(rep(0,dim(training)[1]))
SOplus[iO]=1

tableO=cbind(table.arbre,SOplus)
N=dim(training)[1]
samp=sample(1:N,floor(N/2))
```

La fonction *rpart* du logiciel R produit une multitude d'arbres de décision sous plusieurs contraintes :

- L'arbre doit avoir au minimum 10 branches.
- Lorsque l'arbre ne peut prédire avec certitude l'affectation d'un assuré à la classe 0 ou 1, la priorité est donnée à la classe 0 dans 70% des cas. Dans les 30% autres cas l'assuré obtient la valeur 1. Ces probabilités de priorité de classe permettent d'obtenir la meilleure prédiction possible, sans surestimer ni sous-estimer le nombre de sinistres.

```
Tree=rpart(as.factor(SOplus)~Age+Group1+Bonus+Density+
             Seg.Age.mat+Seg.Bonus.mat+Seg.Density.mat+Seg.Group1.mat,
             data=table0,control=rpart.control(minsplit=10),
             subset=samp,parms=list(prior = c(.7,.3)),method="class")
```

La fonction *prune* de la library *rpart* permet d'obtenir le meilleur arbre de décision, parmi l'ensemble des arbres créés.

```
TreeOplus=prune(Tree,cp=Tree$cptable[which.min(Tree$cptable[,4]),1])
#Tree$cptable = taux d'erreur pour chaque arbre #cp = obtenir l'arbre optimal
```

L'arbre optimal peut être analysé grâce à la fonction *summary*. Les variables `Age` et `Bonus` ainsi que leur segmentation sont les variables explicatives les plus utilisées pour réaliser l'arbre de classification.

```
Call:
rpart(formula = SOplus ~ Age + Group1 + Bonus + Density + Adind +
    Seg.Age.mat + Seg.Bonus.mat + Seg.Density.mat + Seg.Group1.mat,
    data = table0, subset = samp, method = "class", parms = list(prior = c(0.7,
        0.3)), control = rpart.control(minsplit = 5, cp = 0))
  n= 50010

           CP nsplit rel error    xerror       xstd
1  0.039832768      0 1.0000000 1.0000000 0.01204697
2  0.015332996      2 0.9203345 0.9203345 0.01043378
3  0.010489221      4 0.8896685 0.8918858 0.01076721
4  0.003816104      6 0.8686900 0.8795886 0.01038789
5  0.002695958      8 0.8610578 0.8744361 0.01030203
6  0.002483979     11 0.8529699 0.8732344 0.01025508
7  0.002106374     13 0.8480020 0.8722157 0.01020353
8  0.001467615     14 0.8458956 0.8645802 0.01010736
9  0.001259048     17 0.8414928 0.8621919 0.01008868
10 0.001064127     20 0.8377156 0.8614121 0.01012366
11 0.001003308     23 0.8345232 0.8611291 0.01012066

Variable importance
          Bonus    Seg.Bonus.mat              Age     Seg.Age.mat Seg.Density.mat
             33               32               10              10               8
        Density          Group1  Seg.Group1.mat
              4               2               2
```

Le premier nœud de l'arbre regroupe l'ensemble des assurés présents dans la base d'apprentissage. Cette population est séparée en deux sous-population filles. L'affectation à une sous-population dépend des caractéristiques des individus.

```
Node number 1: 50010 observations,    complexity param=0.03983277
  predicted class=0  expected loss=0.3  P(node) =1
    class counts: 43954  6056
   probabilities: 0.700 0.300
  left son=2 (28587 obs) right son=3 (21423 obs)
  Primary splits:
      Bonus           < -15      to the left,  improve=1777.6780, (0 missing)
      Seg.Bonus.mat   < 2.5      to the left,  improve=1777.6780, (0 missing)
      Seg.Age.mat     < 8.5      to the left,  improve=1101.8730, (0 missing)
      Age             < 29.5     to the right, improve=1101.8730, (0 missing)
      Seg.Density.mat < 8.5      to the left,  improve= 624.8524, (0 missing)
  Surrogate splits:
      Seg.Bonus.mat < 2.5      to the left,  agree=1.000, adj=1.000, (0 split)
      Age           < 30.5     to the right, agree=0.676, adj=0.243, (0 split)
      Seg.Age.mat   < 8.5      to the left,  agree=0.675, adj=0.242, (0 split)
```

	Vecteur de prédiction	
Variable à prédire	39 535	4 255
	3 926	2 295

TABLE 4.11 – *Table de performance du nombre de sinistres matériels pour le modèle d'arbres de décisions*

	Vecteur de prédiction	
Variable à prédire	45 615	2 023
	1 914	310

TABLE 4.12 – *Table de performance du nombre de sinistres corporels pour le modèle d'arbres de décisions*

Le taux de performance des sinistres matériels est de 35% des cas, tandis que celui des sinistres corporels n'est que de 13%. Le taux d'erreur global est de 16%.

L'arbre de classification optimal utilisé pour la prédiction de sinistres matériels est plus riche, et permet de classer les individus de la base d'apprentissage avec une meilleure précision, ce qui explique la différence de bonne prédiction entre les deux variables à expliquer.

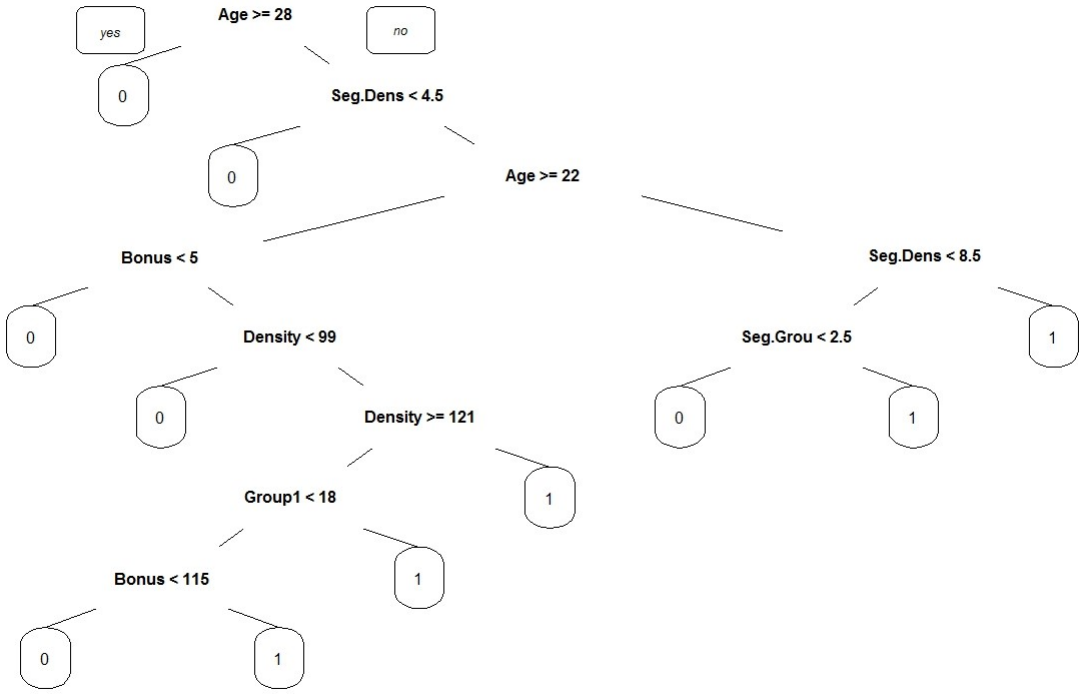

FIGURE 4.8 – *Arbre de classification du nombre de sinistres corporels*

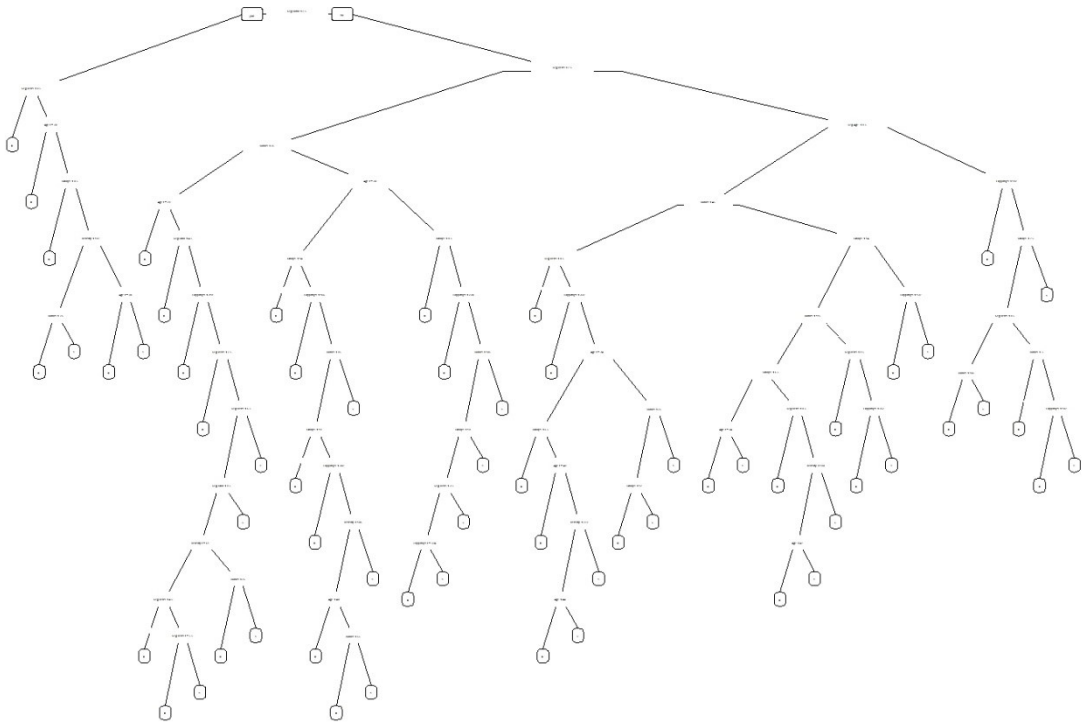

FIGURE 4.9 – *Arbre de classification du nombre de sinistres matériels*

4.2.6 Le boosting

Ce modèle très performant selon Arthur CHARPENTIER, consiste à booster la performance d'algorithmes de prédiction appelés les classifieurs. L'utilisateur a le choix entre plusieurs classifieurs : fonctions linéaires, séparateurs non-linéaires ou arbres de décisions. Parmi les classifieurs sélectionnés, certains sont dit faibles, c'est-à-dire qu'ils impactent peu la variable à expliquer.

La performance de chaque classifieur est mesurée par la fonction de perte :

$$L(F(x), y)$$

La fonction de perte dans notre cas est la suivante :

$$L(x, F(x)) \to (y - F(x))^2$$

La dérivée de la fonction de perte est la suivante :

$$\frac{dL\left(y, F(x)\right)}{dx} = -2\left(y - F(x)\right)$$

Elle mesure la différence entre la valeur exacte de la variable cible y, et la valeur obtenue par le classifieur.

Lors de la première étape de l'algorithme de boosting, un vecteur $(\hat{f}^0)_{i=1..n}$ est créé et initialisé par des valeurs initiales précisées dans les paramètres de l'algorithme. Le nombre n représente le nombre d'individus dans la base d'apprentissage. A chaque étape notée m, la dérivée de la fonction de perte est calculée pour chacun des p classifieurs, et est évaluée en $\hat{f}^{m-1}$. Le vecteur obtenu à chaque étape est le suivant :

$$u^m = \left(-\frac{d\left(L\left(\hat{f}^{m-1}, y_i\right)\right)}{df} \right)_{i=1,..n}$$

On obtient p vecteurs u^m. L'algorithme applique la méthode des moindres carrés à ces p vecteurs afin de sélectionner le classifieur qui approche au mieux u^m. Cette méthode consiste à calculer la somme des carrés des écarts entre la valeur réelle et la valeur obtenue.
En notant p^* la fonction sélectionnée, le vecteur obtenu par le meilleur classifieur est le suivant :

$$\hat{u}^m = \left(-\frac{d\left(L\left(\hat{f}^{m-1}, y_i\right)\right)}{df^{p^*}} \right)_{i=1,..n}$$

Le vecteur $\hat{f}^m$ s'écrit :

$$\hat{f}^{m-1} + v \times \hat{u}^m$$

Avec v est un coefficient compris entre 0 et 1.

La prédiction finale calculée par l'algorithme est :

$$\hat{f} = \sum_{k=0}^{m} \hat{f}^k$$

Cet algorithme peut être interprété comme une fonction $\hat{f}$ équivalente à :

$$\hat{f} = \sum_{k=1}^{p} \hat{f}^k$$

Avec $\hat{f}^k$ la somme des vecteurs $v \times \hat{u}^m$ obtenus au cours des m itérations exécutées avec le meilleur classifieur.

Les différents classifieurs utilisés dans le package *mboost* de R peuvent être des fonctions linéaires, des séparateurs non-linéaires, des arbres de décision.

L'algorithme boosting est implémenté sous R de la manière suivante :

```
i0=which(training$Numtppd>=1)

SOplus=c(rep(0,dim(training)[1]))
SOplus[i0]=1

table0=cbind(tableboosting,SOplus)
samp=echantillon(training$Numtppd)
table0=table0[samp,]
SOplus=SOplus[samp]
```

Les variables explicatives sont utilisées dans des classifieurs linéaires et des arbres de décision. La méthode de descente de gradient est appliquée 200 fois, afin d'obtenir une bonne prédiction de la variable binaire, sans sur-estimé ni sous-estimé le nombre de sinistres prédits.

```
boostOplus=mboost(SOplus~btree(Age)+btree(Group1)+btree(Bonus)+bols(Value)+bols(Density),
                  data=table0,control = boost_control(mstop = 200))
```

L'arbre optimal peut être analysé grâce à la fonction *summary* :

```
        Model-based Boosting

Call:
mboost(formula = S01 ~ btree(Age) + btree(Group1) + btree(Bonus) +     bols(Value)
 + bols(Density), data = training1, control = boost_control(mstop = 200))


        Squared Error (Regression)

Loss function: (y - f)^2


Number of boosting iterations: mstop = 200
Step size:  0.1
Offset:  0.1232354
Number of baselearners:  5

Selection frequencies:
   btree(Age) btree(Group1)  btree(Bonus) bols(Density)
        0.320         0.270         0.265         0.145
```

Les classifieurs ici utilisés sont les classifieurs linéaires et les arbres de décisions.

```
Numtppd=predict(boostOplus,table0[-samp,])
n=prediction(Numtppd,SOplus[-samp,])
Numtppd=floor(Numtppd+as.numeric(n))
```

Les performances du modèle sont montrées dans les Tables 4.13 et 4.14.

	Vecteur de prédiction	
Variable à prédire	39 628	4 162
	4 015	2 206

TABLE 4.13 – *Table de performance du nombre de sinistres matériels pour le modèle boosting*

	Vecteur de prédiction	
Variable à prédire	45 574	2 213
	1 898	326

TABLE 4.14 – *Table de performance du nombre de sinistres corporels pour le modèle boosting*

Le taux de performance du modèle boosting est de 35% pour les sinistres matériels et de 13% pour les sinistres corporels. Le taux d'erreur global est de 16%.

4.3 Résultat de la prédiction

Les différents modèles mis en place permettent de prédire de 6 manières différentes le nombre de sinistres par assuré et de 2 manières différentes le coût des sinistres. En croisant les modèles linéaires et les algorithmes d'apprentissage statistique entre eux, la prime pure de chaque assuré peut être obtenue de 9 façons distinctes.

	Modélisation du nombre de sinistres	Modélisation du coût moyen d'un sinistre
Modèle 1	Régression linéaire	Régression linéaire
Modèle 2	Régression linéaire	GLM gamma
Modèle 3	GLM binomial négatif	Régression linéaire
Modèle 4	GLM Poisson	Régression linéaire
Modèle 5	GLM binomial négatif	GLM gamma
Modèle 6	GLM Poisson	GLM gamma
Modèle 7	SVM	GLM gamma
Modèle 8	Arbre de décision	GLM gamma
Modèle 9	Boosting	GLM gamma

TABLE 4.15 – *Récapitulatif des modèles étudiés*

Les performances des modèles linéaires et d'apprentissage statistique ne sont pas comparables entre elles. La seule façon de le faire, est de comparer les ratios S/P obtenus dans le chapitre 5 du projet. Le modèle qui obtiendra le ratio S/P le plus proche de 1 est celui qui prédit au plus juste le coût des assurés. En revanche, pour dégager le meilleur bénéfice, le modèle devra obtenir le ratio S/P le plus faible.

Un modèle est considéré comme performant s'il prédit au mieux, avec le moins d'erreurs possibles, ce que coûtera les assurés au cours de l'année 2011. Il est donc nécessaire de quantifier cette erreur pour comparer les modèles entre eux.

Pour vérifier la fiabilité des modèles linéaires, il suffit de les projeter sur la base `Training`. C'est-à-dire de prédire les variables à expliquer de la base d'apprentissage `Training` afin de comparer la valeur réelle et la valeur prédite. Le calcul de l'erreur quadratique permet de rendre compte de l'écart entre ces deux variables.

$$Erreur\ quadratique = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Le modèle qui obtient l'erreur quadratique la plus faible dans la prédiction du tarif est le modèle 2 régression linéaire - GLM gamma, suivi de près par le modèle 6 GLM Poisson - GLM gamma. Ce sont donc a priori les modèles qui permettront d'obtenir le meilleur ratio S/P dans la deuxième partie du projet. Le modèle le moins précis est le modèle régression linéaire - régression linéaire.

Le SVM est dominé sur les 4 critères de la table 4.17 par les arbres de décision et l'algorithme de boosting. Les seules indications de ce tableau permettent a priori de classer en dernier le SVM sur les performances, et de classer en premier les arbres de décision en se basant sur les mêmes critères.

	Variables à expliquer	Erreur quadratique
Modèle 1	Nombre de sinistres : régression linéaire Coût moyen : régression linéaire	Matériel : 183 425,05 Corporel : 439 906,09
Modèle 2	Nombre de sinistres : régression linéaire Coût moyen : GLM gamma	Matériel : 183 253,05 Corporel : 411 162,09
Modèle 3	Nombre de sinistres : GLM binomial négatif Coût moyen : régression linéaire	Matériel : 183 421,23 Corporel : 439 901,01
Modèle 4	Nombre de sinistres : GLM Poisson Coût moyen : régression linéaire	Matériel : 183 421,08 Corporel : 439 900,44
Modèle 5	Nombre de sinistres : GLM binomial négatif Coût moyen : GLM gamma	Matériel : 183 249,08 Corporel : 411 156,44
Modèle 6	Nombre de sinistres : GLM Poisson Coût moyen : GLM gamma	Matériel : 183 249,23 Corporel : 411 157,01

TABLE 4.16 – *Erreur quadratique selon le modèle*

Algorithme	Fiabilité	Temps de calcul	Taux de performance	Taux d'erreur globale
Réseaux de neurones	Non	/	/	/
SVM	Oui	-	18%	20%
Arbres de décisions	Oui	+	35%	16%
Boosting	Oui	-	35%	16%

TABLE 4.17 – *Performance des différents modèles d'apprentissage statistique*

Une analyse des vecteurs de sinistres, qui sont calculés à partir de l'ensemble des vecteurs de prédiction, est nécessaire pour départager ces trois méthodes.

Les Figures 4.10, 4.11 et 4.12 sont construits sous la forme suivante : pour chaque nombre de sinistres x compris entre 0 et 7 en abscisse, un calcul de la somme cumulée des sinistres est effectué sur l'ensemble des assurés ayant exactement x sinistres. Les barres noires correspondent aux sinistres répertoriés dans la base `Training`, les autres correspondent aux sinistres prédits. Les valeurs obtenues pour $x = 0$ représentent la somme des fausses prédictions des algorithmes sur l'ensemble des assurés qui ont 0 sinistres dans la base `Training`.

Le faible taux de performance des algorithmes explique les résultats obtenus lors de la prédiction des sinistres corporels : 87% des assurés au minimum ont une prédiction fausse, et seuls 5% des assurés de la base `Training` ont au moins un sinistre corporel. Les mauvaises prédictions s'accumulent donc sur l'ensemble des assurés n'ayant aucun accident.

Les Figures 4.10, 4.11 et 4.12 confirment que le SVM peut être considéré comme l'algorithme le moins performant. La majeure partie de ses prédictions sont fausses, de plus la prédiction sur les assurés les plus risqués est inexistante.
Une analyse similaire permet d'affirmer que le boosting est l'algorithme qui réalise les meilleures prédictions.

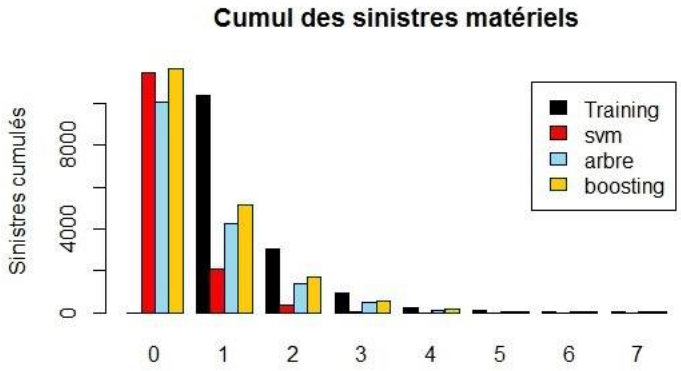

FIGURE 4.10 – *Cumul des sinistres matériels (de 0 à 7 sinistres)*

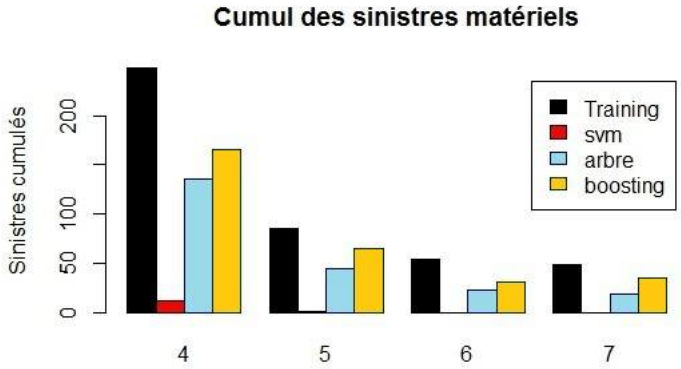

FIGURE 4.11 – *Cumul des sinistres matériels (de 4 à 7 sinistres)*

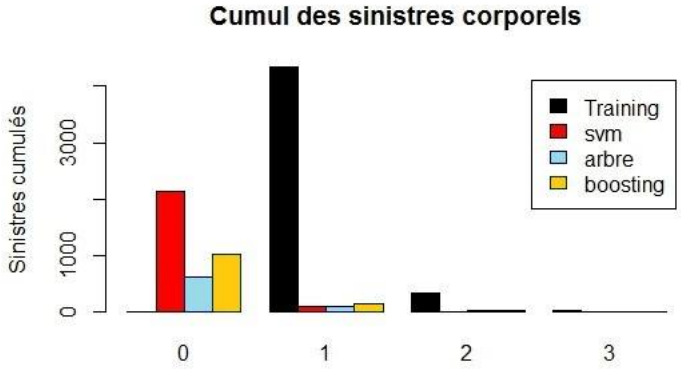

FIGURE 4.12 – *Cumul des sinistres corporels*

4.4 Répartition des sinistres graves

Les sinistres graves sont les sinistres matériels dont le montant est supérieur à 4 000 € et les sinistres corporels dont le montant est supérieur à 30 000 €[3]. Ils n'ont pas été simulés par une loi : il a été décidé de répartir le montant total de ces sinistres entre tous les assurés. C'est-à-dire que chaque assuré paie une part des sinistres graves.

On considère que tous les modèles gèrent les sinistres graves de cette façon.

Le montant total des sinistres graves est composé du montant des sinistres graves matériels (373 605 €) et des sinistres graves corporels (913 811 €) : les 28 908 assurés de la base de test se répartissent les 1 287 416 €. Chaque assuré voit sa prime pure prédite augmentée de 44,53 €.

3. D'après la Théorie des Valeurs Extrêmes page 29.

Chapitre 5

Analyses d'impacts sur le marché

Le travail effectué dans la partie précédente a consisté à prédire les primes pures des assurés de la base `Pricing` par différents modèles. Il est maintenant nécessaire de vérifier que les modèles performants permettent à l'assureur d'exercer une activité viable.

L'indicateur de viabilité d'un assureur est son ratio S/P (voir page 9). Il est défini ainsi :

$$\frac{S}{P} = \frac{\text{Montant total des sinistres à la fin de l'exercice}}{\text{Montant total des primes perçues}}$$

Les primes perçues correspondent aux 9 différents vecteurs obtenus lors de la tarification du contrat au 1er Janvier 2011, tandis que le montant total des sinistres est une variable observée à la fin de l'exercice de l'assureur, soit le 31 Décembre 2011.

Nous pensions parvenir à récupérer cette donnée auprès d'Arthur CHARPENTIER, malheureusement notre requête n'a pas abouti. La base `Pricing` et les vecteurs simulés jusqu'ici ne permettaient donc pas de répondre à la deuxième problématique du sujet.

Nous avons résolu le problème en divisant la base `Training` en deux parties :

- 2/3 des observations servent de nouvelle base d'apprentissage, soit 57 816 observations.
- 1/3 des observations servent de nouvelle base de test. Celle-ci contient 28 908 observations et remplace la base test `Pricing`.

La base d'apprentissage se voit donc fortement diminuée. Pour pallier cette perte, il a été décidé de simuler le marché 100 fois. Chaque fois, le 2/3 d'observations servant de base d'apprentissage et les 1/3 d'observations servant de base de test sont tirés au hasard. La conclusion du marché est toujours la même : les ratios S/P des assureurs sont classés dans le même ordre. Le choix des observations constituant chaque base n'impacte donc pas le résultat final.

Les deux variables nécessaires au calcul du ratio sont à présent disponibles, puisque le montant des sinistres des assurés de la nouvelle base test est disponible au 31 Décembre 2011.

5.1 Analyses des primes

Dans la suite du projet, chaque méthode de tarification est confondue avec un assureur afin d'aborder la problématique uniquement d'un point de vue économique.

Le tracé des diagrammes en boite (ou *boxplot*) permet de visualiser la répartition des primes pures de chaque assureur. Les assureurs qui proposent les primes les plus élevées sont les assureurs A1 et A2. Ils proposent sensiblement les mêmes primes.
Les primes les plus faibles sont proposées par les assureurs A3, A4 et A5.

Assureur	Méthode de tarification
Assureur 1	Régression linéaire - Régression linéaire
Assureur 2	Régression linéaire - GLM gamma
Assureur 3	GLM binomial négatif - Régression linéaire
Assureur 4	GLM Poisson - Régression linéaire
Assureur 5	GLM binomial négatif - GLM gamma
Assureur 6	GLM Poisson - GLM gamma
Assureur 7	Arbre de décision - GLM gamma
Assureur 8	Boosting - GLM gamma
Assureur 9	SVM - GLM gamma

TABLE 5.1 – *Liste des assureurs et les modèles utilisés*

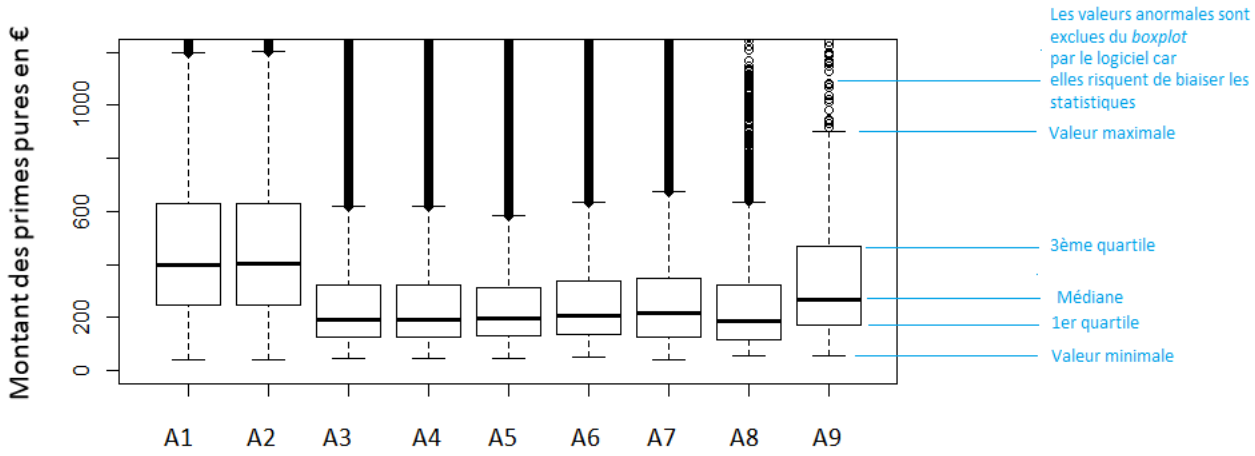

FIGURE 5.1 – *Diagramme en boite des primes pures par assureur*

5.2 Simulation du marché

La mise en concurrence des 9 assureurs sur un marché permet d'observer lesquels proposent le tarif le plus juste. Les primes des assurés qui décident de souscrire un contrat chez chaque assureur sont comparées au coût de leurs sinistres via le calcul du ratio S/P.

La simulation se fait sur le logiciel R : les différents vecteurs de tarif des assureurs sont concaténés dans un même tableau appelé « *Marché* ».

La lecture du tableau en colonne permet de visualiser les différentes primes pures proposées par un assureur à chacun de ses clients, tandis que la lecture en ligne permet de visualiser les différentes primes pures proposées à chacun des clients. La Figure 5.2 présente les 10 premières lignes d'un marché.

La règle de choix des assurés potentiels est la suivante : ils souscrivent un contrat chez l'assureur qui leur propose le tarif le moins cher. Cela revient à sélectionner le minimum sur chaque ligne. La fonction *apply* du logiciel R permet d'inscrire le tarif retenu par chaque client dans une nouvelle colonne *Minimum*. Ce tarif est relié à l'assureur correspondant dans une nouvelle colonne *Répartition*.

Dans la Figure 5.3, le tarif retenu par le client n°1 est 387,93 €. Il souscrit donc son contrat chez l'assureur 7.

5.2.1 Part de marché de l'assureur

La part de marché de chaque assureur se calcule en divisant le nombre de clients qui ont souscrit un contrat dans sa société d'assurance par le nombre total de clients, à savoir 28 908. Le résultat est exprimé en %.

La Figure 5.5 montre que l'assureur 2 et l'assureur 5 détiennent la part de marché la plus importante. Ils

	A1	A2	A3	A4	A5	A6	A7	A8	A9
1	951.45717	948.70517	934.41990	934.40850	895.12432	1033.01746	387.93967	663.75222	699.41986
2	477.25702	498.59634	356.49766	356.49793	284.08396	289.82152	171.55603	231.96571	661.54492
3	1264.41011	1304.57271	1698.79796	1698.79010	874.74498	1036.31197	278.70513	822.32875	619.60774
4	231.69478	228.19833	173.56763	173.56743	144.91685	157.28390	91.83804	73.85274	81.63051
5	1209.31719	1187.14437	827.26744	827.26547	729.97504	811.86333	276.17935	201.56538	619.60774
6	231.54264	245.59058	187.44202	187.44275	188.63256	210.43044	106.06337	98.76208	214.94964
7	221.57179	222.61577	126.26413	126.26432	130.21104	133.90927	111.40315	79.00060	111.97344
8	395.39825	392.24924	298.77656	298.77667	200.78950	217.02266	104.69559	76.97072	81.63051
9	652.03658	645.45301	238.28277	238.28303	229.48183	246.45392	134.76266	120.83168	214.94964
10	949.88298	930.54047	819.71171	819.70399	428.07171	447.92356	867.11549	1688.81264	353.11411

FIGURE 5.2 – *Marché simulé*

	A1	A2	A3	A4	A5	A6	A7	A8	A9	Minimum	Répartition
1	951.45717	948.70517	934.41990	934.40850	895.12432	1033.01746	387.93967	663.75222	699.41986	387.93967	7
2	477.25702	498.59634	356.49766	356.49793	284.08396	289.82152	171.55603	231.96571	661.54492	171.55603	7
3	1264.41011	1304.57271	1698.79796	1698.79010	874.74498	1036.31197	278.70513	822.32875	619.60774	278.70513	7
4	231.69478	228.19833	173.56763	173.56743	144.91685	157.28390	91.83804	73.85274	81.63051	73.85274	8
5	1209.31719	1187.14437	827.26744	827.26547	729.97504	811.86333	276.17935	201.56538	619.60774	201.56538	8
6	231.54264	245.59058	187.44202	187.44275	188.63256	210.43044	106.06337	98.76208	214.94964	98.76208	8
7	221.57179	222.61577	126.26413	126.26432	130.21104	133.90927	111.40315	79.00060	111.97344	79.00060	8
8	395.39825	392.24924	298.77656	298.77667	200.78950	217.02266	104.69559	76.97072	81.63051	76.97072	8
9	652.03658	645.45301	238.28277	238.28303	229.48183	246.45392	134.76266	120.83168	214.94964	120.83168	8
10	949.88298	930.54047	819.71171	819.70399	428.07171	447.92356	867.11549	1688.81264	353.11411	353.11411	9

FIGURE 5.3 – *Répartition du marché*

	A1	A2	A3	A4	A5	A6	A7	A8	A9	Minimum	Répartition	Part de marché
1	951.45717	948.70517	934.41990	934.40850	895.12432	1033.01746	387.93967	663.75222	699.41986	387.93967	7	15.4799140
2	477.25702	498.59634	356.49766	356.49793	284.08396	289.82152	171.55603	231.96571	661.54492	171.55603	7	15.4799140
3	1264.41011	1304.57271	1698.79796	1698.79010	874.74498	1036.31197	278.70513	822.32875	619.60774	278.70513	7	15.4799140
4	231.69478	228.19833	173.56763	173.56743	144.91685	157.28390	91.83804	73.85274	81.63051	73.85274	8	11.0864041
5	1209.31719	1187.14437	827.26744	827.26547	729.97504	811.86333	276.17935	201.56538	619.60774	201.56538	8	11.0864041
6	231.54264	245.59058	187.44202	187.44275	188.63256	210.43044	106.06337	98.76208	214.94964	98.76208	8	11.0864041
7	221.57179	222.61577	126.26413	126.26432	130.21104	133.90927	111.40315	79.00060	111.97344	79.00060	8	11.0864041
8	395.39825	392.24924	298.77656	298.77667	200.78950	217.02266	104.69559	76.97072	81.63051	76.97072	8	11.0864041
9	652.03658	645.45301	238.28277	238.28303	229.48183	246.45392	134.76266	120.83168	214.94964	120.83168	8	11.0864041
10	949.88298	930.54047	819.71171	819.70399	428.07171	447.92356	867.11549	1688.81264	353.11411	353.11411	9	0.8503568

FIGURE 5.4 – *Part du marché*

concentrent à eux deux près de la moitié du marché. A l'inverse, les assureurs 1, 6 et 9 n'attirent à eux trois que 3% environ des clients. Aucun assureur n'est mis de côté : chacun attire au minimum 220 assurés.

L'attractivité ainsi exercée auprès des 28 908 clients par les différents assureurs doit être complétée par une vérification de la viabilité des tarifs proposés. Il ne s'agit pas qu'un grand portefeuille, constitué de trop de mauvais clients, vienne obérer l'équilibre financier de l'assureur et conduise à une activité déficitaire.

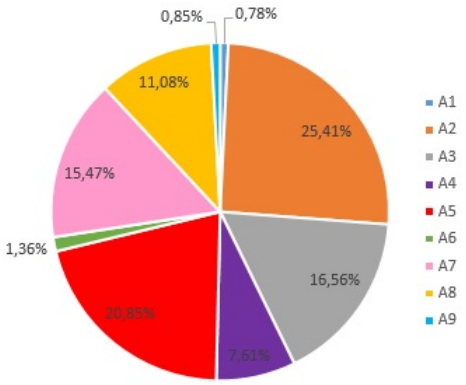

FIGURE 5.5 – *Part du marché*

5.2.2 Ratio S/P et bénéfice de l'assureur

L'activité de l'assureur est viable si son ratio S/P est inférieur à 1, prouvant ainsi que le montant total des primes qu'il a encaissées a au moins couvert l'ensemble des sinistres auxquels il a dû faire face. Il se calcule en divisant le montant total des sinistres auxquels l'assureur à fait face au cours de l'année 2011 par le montant total des primes encaissées au 1er Janvier de cette année civile. Le bénéfice de l'assureur correspond à la différence entre ces deux grandeurs : $P - S$.

	Nombre d assuré	Pourcentage	Prime max	Prime min	Prime moy	S/P	Benef
A1	226.0424	0.7819372	1992.3194	41.27288	255.0442	1.8909546	-200876.1
A2	7346.3787	25.4129606	2043.9512	41.64137	233.5944	0.9134533	21224.4
A3	4787.8610	16.5624084	1244.1753	47.25902	156.5446	0.4500897	2976550.5
A4	2200.3817	7.6116704	1160.8492	50.47092	241.0148	0.7501665	789212.4
A5	6028.2688	20.8532890	1975.2568	47.95981	256.6889	0.9819805	271324.2
A6	393.4551	1.3610595	613.1554	51.53233	155.6082	0.7479532	139578.7
A7	4474.9335	15.4799140	1233.2768	38.61714	136.92044	0.9947625	28380.8
A8	3204.8577	11.0864041	1059.4601	58.05649	138.24458	0.9003615	780616.0
A9	245.8211	0.8503568	874.3356	54.99457	140.04470	1.2612085	-786797.3

FIGURE 5.6 – *Résultat du marché*

L'assureur 1 et 9, qui détiennent la part de marché la plus faible, réalisent une perte. Le peu de clients qu'ils attirent ne paie pas des primes suffisamment importantes par rapport au montant de leurs sinistres. Leur modèle de tarification ne résiste pas à la concurrence.

En revanche, les autres assureurs font un bon résultat. Leur ratio S/P est inférieur à 1, ce qui signifie que les primes encaissées couvrent les indemnisations versées.

L'assureur 3 obtient le meilleur ratio : 45%. Son modèle de tarification binomial négatif - régression linéaire est donc le meilleur. Son important bénéfice révèle néanmoins une surestimation des coûts : les assurés paient plus que nécessaire. Les assureurs 4 et 6 obtiennent un ratio S/P d'environ 75%, celui des assureurs 2 et 8 est d'environ 90% tandis que celui des assureurs 5 et 7 est proche de 1.

Conclusion

Le cycle de production d'un assureur est inversé : l'estimation des coûts du produit d'assurance (montant des sinistres à indemniser) est donc indispensable afin de proposer un tarif à l'assuré. Cette estimation doit être précise afin d'assurer la viabilité de l'activité de l'assureur. Une sous-estimation des coûts du produit entraîne un déséquilibre entre les recettes (primes encaissées) et les dépenses (indemnisation des sinistres), et conduit l'assureur à une activité déficitaire, et donc à terme, à une faillite certaine. Il s'agit donc d'obtenir une bonne adéquation entre la sinistralité et les primes payées par les assurés.

Le cœur de ce bureau d'étude a consisté à tarifer un contrat d'assurance automobile RC à l'aide de différents modèles informatiques, plus ou moins performants. Nous sommes arrivés à la conclusion que les deux modèles de tarification les plus performants (c'est-à-dire qui prédisent avec une erreur minimale) sont le modèle régression linéaire - gamma, utilisé par l'assureur 2 et le modèle boosting - gamma, utilisé par l'assureur 8.

La bonne performance de ces modèles ne garantissant ni l'attractivité de l'assureur vis-à-vis de ses clients potentiels, ni la viabilité de son activité, les modèles ont été comparés d'un point de vue économique. Une simulation de marché a ainsi permis de mettre en concurrence les 9 modèles de tarification. Il était prévu de simuler le marché de deux manières différentes :

- L'assuré choisit l'assureur le moins cher dans un premier marché.
- L'assuré choisit au hasard parmi les assureurs les moins chers dans un second marché.

Cependant, ce second marché n'a pas pu être mis en place : le logiciel R a renvoyé sur nos trois ordinateurs un message d'erreur signalant un problème de mémoire vive, dont la limite aurait été atteinte. Ceci est étonnant puisque le travail effectué par le logiciel ne semblait pas plus important que pour simuler le premier marché.

Même si l'attractivité des clients n'impacte pas directement la viabilité de l'activité de l'assureur, elle permet néanmoins de minimiser les risques de faillite : le montant total des sinistres est reparti entre un nombre d'assurés plus important. La mutualisation des risques est ainsi accrue. On peut retenir que l'assureur 2 se classe en tête en occupant 25% du marché, alors que l'assureur 8 se classe à la cinquième place avec 11% du marché. Les assureurs 1,6 et 9 ne captent à eux trois que 3% du marché.

L'indicateur de viabilité de l'activité de l'assureur est le ratio S/P. Les meilleurs modèles de tarification sont dans l'ordre croissant des ratios S/P :

- Le modèle binomial négatif - régression linéaire (Assureur 3)
- Le modèle Poisson - régression linéaire (Assureur 4)
- Le modèle Poisson - gamma (Assureur 6)
- Le modèle boosting - gamma (Assureur 8)
- Le modèle régression linéaire - gamma (Assureur 2)
- Le modèle binomial négatif - gamma (Assureur 5)
- Le modèle arbre de décision - gamma (Assureur 7)
- Le modèle SVM - gamma (Assureur 9)
- Le modèle régression linéaire - régression Linéaire (Assureur 1)

Les modèles les plus performants (Assureurs 2 et 8) se retrouvent pourtant classé $4^{\text{ème}}$ et $5^{\text{ème}}$ dans notre marché.

L'assureur 3, qui utilise le modèle binomial négatif - régression linéaire, réalise le meilleur ratio S/P : 45% et attire 16% des nouveaux clients. Il est donc le meilleur des modèles étudiés dans ce projet, suivi des modèles Poisson - régression linéaire (Assureur 4) et Poisson - gamma (Assureur 6), bien qu'ils ne fassent pas partie des modèles les plus performants en termes de prédiction.

En revanche, les modèles SVM - gamma (Assureur 9) et régression linéaire - régression linéaire (Assureur 1) ne sont pas compétitifs. Non seulement ils n'occupent qu'une très faible part du marché, mais qui plus est ils présentent des ratio S/P tellement élevés, qu'ils réalisent une perte. Rappelons que seule la prime pure a été prise en compte et que l'application d'une marge commerciale permettrait de diminuer sensiblement le ratio.

Les résultats obtenus sont proches de ceux d'Arthur CHARPENTIER : à l'issue du colloque, il a révélé que les deux modèles les plus performants étaient le modèle composé d'un GLM à la fois pour le nombre et le coût moyen d'un sinistre (nos Assureurs 5 et 6) et le modèle boosting (notre assureur 8). Nos deux meilleurs modèles prédisent bien le nombre de sinistres par un GLM, mais prédisent cependant le coût moyen par régression linéaire. Le modèle classé troisième est bien composé de deux prédictions par GLM. Le modèle boosting est ici classé quatrième.

Cette différence s'explique peut-être par notre choix d'écarter la variable `Group2`, plus vraisemblablement notre décision de supprimer les valeurs aberrantes. La raison la plus évidente des distorsions observées s'explique par le changement de base de test que nous avons dû opérer du fait de l'absence des variables observées de la base `Pricing`.

Un approfondissement possible de ce projet consisterait à simuler le marché en 2012 afin de vérifier si le résultat est le même. L'estimation des coûts serait cette fois-ci basée sur la sinistralité de l'année 2011 des assurés.

Ce bureau d'étude a constitué pour nous trois une première expérience professionnelle très enrichissante. Nous avons vu à travers ces travaux les techniques de tarification d'un contrat d'assurance, compétence indispensable pour un actuaire. Par ailleurs, ce projet a permis d'appliquer la globalité du programme informatique vu au cours des deux premières années de formation sur le logiciel R.

Annexe

Rappel sur les lois de probabilités

Expression de la loi de Poisson :

$$P[X = k] = \frac{\lambda^k}{k!} e^{-\lambda} \qquad avec\ \lambda\ positif.$$

Expression de la loi binomiale-négative :

$$P[X = k] = \frac{\Gamma(k+r)}{\Gamma(r)k!} p^r (1-p)^k \qquad avec \quad \Gamma(r) = \int_0^{+\infty} t^{r-1} e^{-t} dt$$

Expression de la densité de la loi gamma :

$$f(x) = x^{k-1} \frac{e^{\frac{-x}{\theta}}}{\Gamma(k)\theta^k}$$

Expression de la densité d'une loi normale :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

avec μ la moyenne de la variable aléatoire et σ^2 la variance de la variable aléatoire.

Loi log-normale :

Si X suit une loi normale de paramètre μ et σ, alors $Y = e^X$ suit une loi log-normale.

Bibliographie

[1] Pierre Ailliot. Cours modèles linéaires. Euria, 2015.

[2] Pierre Ailliot. Cours théorie des valeurs extrêmes. Euria, 2015.

[3] Arthur Charpentier. Statistique de l'assurance. Université de Rennes 1 et Université de Montréal, 2010.

[4] Arthur Charpentier and Michel Denuit. *Mathématiques de l'assurance non-vie*. Economica, 2005.

[5] Arthur Charpentier and Christophe Dutang. L'actuariat avec R. dec 2012.

[6] Brice Franke. Cours processus stochastiques et assurance non vie. Euria, 2016.

[7] William Géhin. Modélisation des queues de distribution des rendements des actifs financiers. Application à la mesure du risque de marché et à la détermination de stratégies d'investissement. Master's thesis, Euria, 2011.

[8] Benjamin Hofner, Andreas Mayr, Nikolay Robinzonov, and Matthias Schmid. Model-based Boosting in R : A hand-on tutorial using the R package mboost. 2014.

[9] Micheline Kamber, Jian Pei, and Jiawei Han. *Data mining : Concepts et techniques*. Morgan Kaufmann, 3rd edition, 2011.

[10] Antoine Paglia and Martial Phelippe-Guinvarc'h. Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique, 2011.

[11] Maxime Richard. Régression régularisée utilisée pour la tarification d'une offre d'assurance bicycles et tricycles à moteur. Master's thesis, Université de Paris-Dauphine, 2011.

[12] Franck Vermet. Cours apprentissage statistique. Euria, 2015.