Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 17/03/2021

Par : **Julie THILL**

Titre : **Modèle de rétention suite à remplacement sur le périmètre Moto**

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière
Christian-Yann ROBERT

Entreprise : RUN Services pour AXA France
Nom : Alix RAMBAUD
Signature :

Membres présents du jury de l'Institut des Actuaires

Etienne FLICHY

Florence PICARD

Jérôme SCHAEFFER

Directeur du mémoire en entreprise :

Nom : Alix RAMBAUD
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)
Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Avant propos

La lecture d'un mémoire prend beaucoup de temps et peut paraître fastidieuse. Afin de rendre la lecture de ce mémoire plus ludique, des QRcode sont disponibles dans certaines parties. En scannant ces derniers à l'aide de votre téléphone, vous pourrez m'entendre lire la partie concernée. Vous devez simplement disposer de l'application soundcloud sur votre téléphone et l'enregistrement se lancera automatiquement. Les QRcode sont placés juste en dessous du titre des parties ou chapitres et permettent d'écouter la partie ou le chapitre entièrement.

Cette idée de pouvoir écouter certaines parties du mémoire m'est venue de podcasts que j'ai récemment écoutés. Pouvoir développer ses connaissances sur un sujet en quelques minutes tout en étant dans les transports ou en marchant est quelque chose qui me plaît. C'est pour cela que j'ai tenté de reproduire cette technique pour aborder le sujet de l'Actuariat.

Aussi, je tiens à préciser que pour des raisons de facilité d'écriture et de lecture, j'ai écrit l'intégralité de mon mémoire au masculin et n'ai pas eu recours à l'écriture inclusive.

Résumé

Les motos constituent un marché bien spécifique de l'assurance des véhicules motorisés tant par leurs usages multiples que par la relation les lie à leur conducteur. Le contrat est donc amené à être modifié au gré des envies et des moyens des motards. Cette modification de contrat, aussi appelée remplacement, peut engendrer une hausse ou une baisse de cotisation et potentiellement une résiliation.

L'étude menée consiste à identifier et prédire les résiliations consécutives à un remplacement, c'est à dire qui interviennent 180 jours après cette modification de contrat, afin d'augmenter la rétention des clients et de chiffre d'affaire. La modélisation supervisée de la variable cible binaire a été effectuée à l'aide d'un modèle linéaire généralisé logit. Les mesures de convergence du modèle ainsi que la qualité des prédictions ont permis de valider son utilisation.

La mise en place de mesures tarifaires adaptées ont permis d'améliorer la rétention au moment du remplacement d'environ 3%. Les résultats sont néanmoins à interpréter avec prudence et des axes d'amélioration de l'étude seront proposés.

Mots clés : assurance moto, remplacements, terme, résiliations, modèle linéaire généralisé (GLM), logit.

Abstract

Motorcycles represent a very specific market of motor vehicule insurance both because of their multiple uses and the non-linear relationship between them and their driver. The policy is therefore going to change according to the wishes and means of bikers. This contract modification, also called replacement, can lead to an increase or a decrease in the premium and potentially a termination.

The study consists in identifying and predicting the terminations following a replacement, i.e. which occur 180 days after this contract modification, in order to increase customer retention and turnover. Supervised modeling of the binary target variable was performed using a logit generalized linear model. The model's convergence measures as well as the quality of the predictions validated its use.

The implementation of appropriate pricing measures has improved retention at the time of replacement by around 3%. The results should nevertheless be interpreted with caution and areas for improvement of the study will be suggested.

Key words : motorcycle insurance, replacement, maturity, terminations, generalized linear models (GLM), logit.

Note de synthèse

Contexte de l'étude et problématique

Les motos constituent un marché bien spécifique de l'assurance des véhicules motorisés. Elles sont utilisées par les passionnés qui se retrouvent pour des promenades, par les sportifs sur des circuits ou encore comme simple moyen de transport pour les trajets du quotidien ([10] *Institut de sondage Kantar TNS, 2020*). Les usages sont donc multiples et les risques associés différents. De plus, la relation entre l'individu et sa moto est différente d'un assuré à l'autre et évolue dans le temps au grès de ses envies et de ses moyens. En 2019, les véhicules deux roues représentent 8,2% de l'ensemble du parc de véhicules assurés. AXA se place en 4$^{\text{ème}}$ position dans le classement des assureurs détenant le plus de contrats deux roues et générant le plus de chiffre d'affaire sur cette branche. L'offre complète d'assurance proposée en collaboration avec RUN Services permet de répondre aux besoins de l'ensemble des usagers, qu'ils soient passionnés ou simples utilisateurs. Néanmoins, le marché de l'assurance deux roues est en forte croissance et la concurrence entre les différents acteurs du marché est bien réelle. Le client cherche en effet à bénéficier de la meilleure protection possible au tarif le plus avantageux. Ce climat concurrentiel est d'autant plus visible lors d'une modification de contrat pouvant engendrer une hausse ou une baisse de la prime et donc une potentielle résiliation. C'est en ce sens que l'étude des remplacements, qui correspondent à une modification de contrat, est nécessaire pour avoir accès à un nouveau levier de rétention de clients et donc de chiffre d'affaire.

L'objectif principal de l'étude est de déterminer un modèle de rétention suite à remplacement sur le périmètre moto. Cette étude a pour but d'identifier la structure de ces faits de production, leur fréquence en portefeuille ainsi que la perte de chiffre d'affaire qu'ils engendrent. Elle doit également permettre de déterminer les caractéristiques des remplacements les plus susceptibles de résilier afin de pouvoir prédire et surtout éviter, à l'aide de mesures tarifaires adaptées, les résiliations futures.

Mise en forme et analyse des données

Délai entre le remplacement et la résiliation

Les données utilisées doivent être adaptées au périmètre étudié, pertinentes et exactes. De plus, aucune modification majeure de structure de celles-ci ne doit avoir eu lieu sur la période temporelle sélectionnée. En ce sens, un historique de remplacements de 3 ans, sur la période de 2017 à 2019, et un historique de résiliations de 3 ans et 3 mois, sur la période de 2017 à mars 2020, ont été sélectionnés. Les données de résiliation postérieure à mars 2020 n'ont pas été prises en compte, bien qu'elles fussent disponibles, du fait du contexte économique incertain lié à cette période. Le processus de création et de fiabilisation des bases de données fut long et fastidieux, mais néanmoins nécessaire pour mener à bien ce projet.

Grâce à cette base de données, il est maintenant possible de définir dans quels cas les résiliations sont considérées comme consécutives à un remplacement et dans quels cas elles sont considérées indépendantes du remplacement. Une

étude du délai entre la modification de contrat et la résiliation a donc été menée et a permis de retenir l'hypothèse selon laquelle que la résiliation est consécutive à un remplacement si elle intervient dans les 180 jours qui suivent cette modification de contrat. Les remplacements non résiliés qui ne disposent pas de cette fenêtre d'observation de 180 jours ont été supprimés de la base pour conserver l'homogénéité de cette dernière. La base finale des remplacements contient 167 929 observations dont 9 616 d'entre elles sont des résiliations suite à modification de contrat. Les différents motifs de remplacement et de résiliation ainsi que la répartition des observations selon ces motifs sont illustrés dans les tableaux ci-dessous.

Motifs de remplacement	Proportion de remplacements pour chaque motif
1/ changement de véhicule	36,8%
9/ modification tarifaire	16,5%
3/ diminution de garanties	15,3%
4/ changement de zone-cdp	12,9%
2/ extension de garanties	6,4%
5/ changement de sitma/usage/csp	6,1%
7/ ajout d'une clause	3,8%
6/ changement de fractionnement	1,5%
8/ retrait d'une clause	0,7%

Motifs de résiliation	Proportion de résiliations suite à remplacement pour chaque motif
Vente	72,0%
Hamon	15,7%
Loi Chatel	2,8%
Remplace par	2,7%
Echeance	2,2%
Autres cas	1,4%
Sinistre	1,1%
Suite suspension	1,0%
Changement de situation	0,4%
Décès	0,4%
Perte Totale	0,3%
Refus de majoration	0,0%
Liquidation judiciaire	0,0%

Répartition des remplacements et des résiliations suite à remplacement par motifs

Source : Portefeuille AXA (Excel)

Création d'indicateurs et discrétisation des variables continues

En ce qui concerne les variables de la base de données, certains indicateurs ont été créés pour potentiellement augmenter le caractère explicatif du modèle. Les principaux indicateurs sont les suivants :

— **Évolution du coefficient technique** : le coefficient technique, appelé CT, correspond au ratio entre la prime payée et la prime à l'affaire nouvelle. Il permet d'identifier un éventuel rabais appliqué à un moment de la vie du contrat mais également l'évolution du tarif selon les années. L'objectif est de ramener le coefficient technique à 1 lorsque ce dernier est inférieur à cette valeur. En effet, à l'affaire nouvelle, on a

$$CT = \frac{Prime\_payée}{Prime\_affaire\_nouvelle} = \frac{Prime\_affaire\_nouvelle}{Prime\_affaire\_nouvelle} = 1$$

— **Écart de cotisation** : cette variable indique simplement la différence entre les cotisations après et avant remplacement. Elle est exprimée en euros. Une valeur négative indique que la cotisation avant est plus élevée que celle après et inversement.

— **Mois d'effet du remplacement** : cet indicateur peut permettre d'identifier une saisonnalité des remplacements. En effet, la pratique de la moto est plus courante pendant les mois d'été qui sont synonymes de beau temps, que pendant les mois d'hiver. Les clients peuvent donc potentiellement effectuer un remplacement au début de l'été pour augmenter leurs garanties et effectuer un deuxième remplacement à la fin de l'été pour diminuer leurs garanties, la moto restant stationnée dans un garage.

— **Nombre de remplacements** : cet indicateur compte le nombre de remplacements effectués sur le contrat pendant la période de 2017 à 2019. Il a été identifié que certains contrats avaient plus de 10 remplacements sur 3 ans ce qui est conséquent.

Afin de pouvoir obtenir un coefficient explicatif par modalité, une discrétisation des variables continues est nécessaire. Cette technique permet de regrouper les modalités des variables continues en différentes classes de valeurs représentées par un intervalle. L'outil utilisé dans le cadre de cette étude pour effectuer cette répartition est la méthode des quantiles. Une fois le nombre de quantiles déterminé, la variable continue est répartie selon les classes créées par ces quantiles. Si le taux de résiliation est sensiblement le même entre deux classes, alors celles-ci peuvent être regroupées.

Etude des corrélations entre les variables

L'étude de la corrélation entre deux variables est différente selon que celles-ci soient quantitatives ou qualitatives. La corrélation entre les variables quantitatives ([11]*RAKOTOMALALA, 2017*) est mesurée à l'aide du coefficient

de Pearson qui vaut 1 si les deux variables sont parfaitement corrélées, -1 si les deux variables sont parfaitement anti-corrélées, et une valeur entre ces deux bornes sinon. La significativité de cette corrélation est déterminée à l'aide du T test qui permet d'indiquer si le coefficient de Pearson est significativement différent de 0. Dans le cadre de ce mémoire, les variables continues ayant été discrétisées, une illustration des corrélations avant discrétisation a été présentée à titre indicatif uniquement. Seules les corrélations entre variables qualitatives seront à retenir par la suite. Elles sont déterminées à l'aide du test d'indépendance de Chi-deux et l'intensité de cette relation est mesurée à l'aide du V de Cramer. Plusieurs corrélations importantes entre les variables ont été identifiées et la suppression d'une d'entre elles s'est révélée nécessaire. Afin de choisir la meilleure variable pour le modèle, un XGBoost ([9] *MELLO, 2020*) a été implémenté sur la base totale pour déterminer l'importance de chacune d'entre elles. La variable corrélée la moins importante d'après l'XGBoost a donc été supprimée pour chaque couple et au total 13 d'entre elles ont été éliminées. Enfin, la corrélation entre les variables qualitatives et la variable cible, la résiliation suite à remplacement, a été calculée. Celles qui sont les moins corrélées avec la variable à prédire peuvent être écartées de l'étude, ce qui est notamment le cas du changement de fractionnement.

Modélisation, mesures de qualité et résultats

La modélisation permet de traduire une situation réelle en langage mathématique. Dans le cadre de cette étude, la variable qui doit être modélisée est binaire, c'est à dire qu'elle ne prend que deux valeurs : 1 si le contrat est résilié dans les 180 jours qui suivent le remplacement, 0 sinon. Pour identifier les caractéristiques des résiliations suite à remplacement et prédire celles à venir, la technique d'apprentissage supervisé est utilisée. Elle permet de séparer la base de données initiale en deux sous-échantillons : le premier est l'échantillon d'apprentissage, sur lequel le modèle va apprendre à reconnaître les résiliations suite à remplacement, et le deuxième est l'échantillon test, sur lequel le modèle va reproduire ce qu'il a appris. Cette séparation en deux sous-échantillons est indispensable afin que le modèle puisse être testé dans des conditions réelles, c'est à dire sans connaître les informations à l'avance.

Le modèle linéaire généralisé

Une technique adaptée pour modéliser la variable cible est le modèle linéaire généralisé logit. Ce modèle a été choisi car il permet de modéliser les variables binaires et est facile à interpréter, contrairement au probit par exemple. La qualité de la modélisation est mesurée par les critères de convergence tels que la déviance, l'Akaike information criterion et le Bayesian information criterion. Ces trois critères doivent être les plus faibles possibles pour que le modèle soit le meilleur possible. L'interprétation de ces mesures de qualité ne peut être effectuée seule, elles doivent être comparées à d'autres modèles.

Modèle GLM logit			
AIC	BIC	Deviance	Loglik
51 690	53 916	59 116	-25 618

Mesures de qualité du modèle GLM logit

Source : Sortie R (mise en forme Excel)

Ensuite, le modèle a été appliqué sur l'échantillon test afin de prédire les remplacements les plus susceptibles de résilier dans les 180 jours. Une probabilité de résiliation a été attribuée à chaque remplacement et une binarisation de celle-ci est nécessaire. Pour ce faire, deux approches ont été utilisées. La première consiste à déterminer le seuil optimal de prédiction en termes de sensibilité, proportion de résiliés correctement identifiés, et spécificité, proportion de non résiliés correctement identifiés. Le seuil obtenu vaut 0,06 ce qui implique que lorsque la probabilité de résiliation est inférieure à 0,06 alors celle-ci est transformée en 0, non résilié, et lorsqu'elle est supérieure ou égale à 0,06 alors elle est transformée en 1, résilié. La deuxième approche consiste à fixer arbitrairement dix seuils, de 0,05 à 0,5 par pas de 0,05. Le tableau ci-dessous indique pour chacun des seuils le nombre de contrats non résiliés correctement prédits (TN), le nombre de contrats résiliés mal prédits (FN), le nombre de contrats non résiliés mal prédits (FP) ainsi que le nombre de contrats résiliés bien prédits (TP). Différentes mesures de performance ont également été calculées pour déterminer quels sont les meilleurs seuils de prédiction et quels sont ceux à écarter. La F1-mesure permet un arbitrage entre la spécificité et la sensibilité : plus elle est élevée, meilleur est le modèle. Ces indicateurs doivent être analysés conjointement. En effet, maximiser la précision et minimiser le taux d'erreur

reviendrait à choisir le seuil de 0,5. Néanmoins, ce seuil permet de prédire presque parfaitement les contrats non résiliés et de capter seulement 0,4% des résiliés, qui correspondent à ceux que l'on souhaite capter dans cette étude.

Seuil	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	Specificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
0,05	20 037	439	11 657	1 453	63,2%	76,8%	11,1%	36,0%	19,4%
0,06	22 010	559	9 684	1 333	69,4%	70,5%	12,1%	30,5%	20,7%
0,10	26 953	951	4 741	941	85,0%	49,7%	16,6%	16,9%	24,8%
0,15	29 499	1 309	2 195	583	93,1%	30,8%	21,0%	10,4%	25,0%
0,20	30 647	1 569	1 047	323	96,7%	17,1%	23,6%	7,8%	19,8%
0,25	31 206	1 708	488	184	98,5%	9,7%	27,4%	6,5%	14,4%
0,30	31 477	1 781	217	111	99,3%	5,9%	33,8%	5,9%	10,0%
0,35	31 592	1 828	102	64	99,7%	3,4%	38,6%	5,7%	6,2%
0,40	31 643	1 860	51	32	99,8%	1,7%	38,6%	5,7%	3,2%
0,45	31 677	1 877	17	15	99,9%	0,8%	46,9%	5,6%	1,6%
0,50	31 685	1 885	9	7	100,0%	0,4%	43,8%	5,6%	0,7%

Prédictions et mesures de performance GLM logit

Source : Sortie R (mise en forme Excel)

Afin de déterminer la capacité de généralisation du modèle GLM et la stabilité de celui-ci, une validation croisée k-fold a été menée. Cette technique itérative permet d'utiliser chacune des observations de la base dans l'échantillon d'apprentissage et dans l'échantillon test. Une validation croisée 10-folds a été menée ce qui signifie que pour chacune des 10 itérations, un modèle linéaire généralisé logit a été implémenté sur des échantillons train et test différents. En moyenne sur les 10 itérations, le seuil optimal vaut 0,6, la spécificité 70,8%, la sensibilité 70,5%, la précision 12,8%, le taux d'erreur 29,2% et la F1-mesure 21,7%. Ces résultats sont très proches de ceux du GLM logit initial ce qui permet de confirmer que le modèle est stable et peut être généralisé.

La procédure step

Une deuxième modélisation, basée sur le GLM logit, a ensuite été menée pour challenger les résultats. La procédure step est un outil de sélection des variables en fonction d'un critère de performance, dans le cadre de ce mémoire, l'AIC. Pas à pas, les variables qui augmentent l'AIC sont retirées du modèle ce qui permet de ne conserver que les variables les plus importantes. Il existe trois méthodes step différentes : forward, backward et stepwise. La méthode backward est celle qui a été implémentée ici. Elle permet, à partir du modèle complet qui contient toutes les variables explicatives, d'éliminer la variable qui fait augmenter l'AIC à chaque itération. Cette direction a été privilégiée car une première sélection des variables a été faite en amont. Les variables prises en entrée par le step sont donc plus susceptibles d'être explicatives du modèle et peu d'entre elles devraient être supprimées. Quatre variables de la base ont été supprimées et ont permis d'améliorer l'AIC de 14 points. Le BIC a également été légèrement amélioré grâce à la procédure step. La déviance quant à elle est restée stable et la log-vraisemblance s'est dégradée.

Modèle step			
AIC	BIC	Deviance	Loglik
51 676	53 373	59 116	-25 665

Mesures de qualité du modèle step

Source : Sortie R (mise en forme Excel)

De même que pour le GLM logit, le step a été appliqué sur les données de test pour déterminer la probabilité de résiliation de chaque nouveau remplacement. La binarisation a été effectuée selon les deux mêmes approches et le seuil optimal en termes de sensibilité et de spécificité vaut également 0.06. Les mesures de performance du modèle step semblent donner les mêmes conclusions que pour le GLM logit à savoir que les meilleures prédictions sont entre 0.05 et 0.20. Les mesures de qualité pour le seuil optimal sont légèrement meilleures pour le step que pour le logit.

Seuil	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	Specificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
0,05	19 969	432	11 725	1 460	63,0%	77,2%	11,1%	36,2%	19,4%
0,06	22 358	579	9 336	1 313	70,5%	69,4%	12,3%	29,5%	20,9%
0,10	26 891	967	4 803	925	84,8%	48,9%	16,1%	17,2%	24,3%
0,15	29 560	1 315	2 134	577	93,3%	30,5%	21,3%	10,3%	25,1%
0,20	30 675	1 575	1 019	317	96,8%	16,8%	23,7%	7,7%	19,6%
0,25	31 235	1 713	459	179	98,6%	9,5%	28,1%	6,5%	14,2%
0,30	31 477	1 782	217	110	99,3%	5,8%	33,6%	6,0%	9,9%
0,35	31 600	1 836	94	56	99,7%	3,0%	37,3%	5,7%	5,5%
0,40	31 646	1 859	48	33	99,8%	1,7%	40,7%	5,7%	3,3%
0,45	31 675	1 873	19	19	99,9%	1,0%	50,0%	5,6%	2,0%
0,50	31 688	1 887	6	5	100,0%	0,3%	45,5%	5,6%	0,5%

Prédictions et mesures de performance step

Source : Sortie R (mise en forme Excel)

La validation croisée n'a pu être implémentée pour le modèle step du fait du temps d'exécution de celle-ci. La procédure step initiale a en effet mis plusieurs heures avant de donner des résultats. De ce fait, le modèle GLM logit est retenu dans le cadre de cette étude pour prédire les résiliations consécutives à un remplacement. Ce choix a également été motivé par le fait que les performances des modèles sont assez similaires et que la capacité de généralisation du modèle GLM logit a été démontrée, contrairement à celle du step.

Impacts concrets de l'étude sur la rétention

L'étude de l'impact de la modélisation sur le chiffre d'affaire est nécessaire pour justifier son intérêt. A partir de la matrice de confusion de la table *Prédictions et mesures de performance GLM logit* présentée plus haut, une comparaison de la rétention sans mesure tarifaire et avec mesure tarifaire a été initiée. Les montants de prime portefeuille et prime de résiliation ont été modifiés pour des raisons de confidentialité. Les proportions ont néanmoins été conservées. Avant résiliations, le chiffre d'affaire s'élève à 7,9M€. Si aucune mesure tarifaire n'est appliquée, ce dernier chute de 1M€ d'après nos prédictions. Afin de limiter les départs et donc augmenter la rétention, des hypothèses tarifaires ont été appliquées. Celles-ci sont déterminées de façon arbitraire et devront être optimisées lors d'une prochaine étude plus large sur le périmètre moto. Une baisse de 5% de la prime est appliquée lorsque les contrats sont prédits comme résiliés. Le taux de conservation des contrats qui ont réellement résilié (*true positive*) est estimé à 75% après cet avantage. Néanmoins, certaines résiliations ne permettront pas de conserver l'assuré, même avec une baisse tarifaire, notamment dans le cas d'une liquidation judiciaire. Au maximum il est possible de conserver 64% des résiliations. Le taux de conservation dans ce cas est donc de 48%.

	EVOLUTIONS DE CHIFFRE D'AFFAIRE		
Seuils	CA avec mesures / CA sans mesure - 1	CA avec mesures / CA avant résiliations - 1	CA sans mesure / CA avant résiliations - 1
0,05	3,2%	-9,7%	-12,5%
0,06	3,1%	-9,8%	-12,5%
0,10	2,5%	-10,2%	-12,5%
0,15	1,7%	-11,0%	-12,5%
0,20	1,0%	-11,6%	-12,5%
0,25	0,6%	-12,0%	-12,5%
0,30	0,4%	-12,1%	-12,5%
0,35	0,2%	-12,3%	-12,5%
0,40	0,1%	-12,4%	-12,5%
0,45	0,1%	-12,4%	-12,5%
0,50	0,0%	-12,4%	-12,5%

Comparaison des évolutions de chiffre d'affaire

Source : Sortie R (mise en forme Excel)

Finalement, la mesure tarifaire appliquée permet de conserver jusqu'à environ 3% du chiffre d'affaire. Néanmoins, ces résultats doivent être analysés avec prudence. En effet, les mesures tarifaires ainsi que le taux de conservation ont été déterminés de façon totalement arbitraire. Une étude de la sensibilité au prix des clients est nécessaire sur le périmètre. De plus, aucun optimum ne semble avoir été identifié dans ce cas avec les paramètres sélectionnés. L'étude a néanmoins permis de comprendre la structure des remplacements et de prédire ceux qui sont les plus susceptibles de résilier dans les 180 jours. La mise en place opérationnelle de cette étude sera effectuée après une analyse plus détaillée de la meilleure mesure tarifaire à mettre en place.

Executive summary

Context of the study and issue

Motorcycles represent a very specific market of motor vehicule insurance. They are used by enthusiasts who get together for rides, athletes on circuits or as a simple means of transport for everyday journeys ([10] *Institut de sondage Kantar TNS, 2020*). The uses are therefore multiple and the associated risks different. Moreover, the relationship between the individual and his two-wheeled vehicule is not linear and evolves over time according to his desires and means. In 2019, two-wheeled vehiculs represent 8,2% of the total insured fleet. AXA is ranked 4th in the ranking of insurers holding the most motorcycle contracts and generating the most turnover in this branch. The insurance offer in collaboration with RUN Service makes it possible tu meet the needs of all users, whether they are enthusiasts or simple users. Nevertheless, the two-wheel insurance market is sharply increasing and competition between the different market players is real. The customer seeks to get the best possible protection at the most advantageous price. This competitive climat is all the more visible when a contract modification may lead to an increase or a decrease in the premium and therefor a potential termination. It is in this sens that the sudy of replacements, which corresponds to a modification of contract, is necessary to have access to a new customer and turnover lever.

The main purpose of the study is to determine a retention model following replacement on the motorcycle perimeter. Until now, motorcycle replacements had not been studied in favor of more ugent tasks. The aim of this study is therefore to identify the structure of these events, their frequency in the portfolio as well as the loss of turnover they generate. It should also make it possible to determine the characteristics of the replacements most likely to terminate in order to be able to predict and, above all, avoid future terminations using suitable pricing measures.

Data formating and analysis

Time between replacement and termination

The data used must be adequate for the selected scope, relevant and accurate. In addition, no major changes in the structure of the data must have taken place over the selected time period. Accordingly, a 3-year replacement history, over 2017 to 2019, and a 3-year and 3-month termination history, over 2017 and March 2020, were selected. Termination data after March 2020 have not been taken into account, althought they were available, due to the uncertain economic environment related to this period. The process of creating and improving the reliability of the databases was long and tedious, but necessary to complete this project.

Thanks to this database, it is now possible to define in which cases terminations are considered as consecutive to a replacement and in which cases they are considered independent of the replacement. A study of the time period between replacement and termination was therefore carried out and conclued that termination is the result of a replacement if it occurs whitin 180 days following this modification of contract. Non-terminated replacements that do not have this 180-day observation window have been removed from the database in order to make it homogeneous. The final replacement database contains 167,929 observations, of which 9,616 are terminations following replacement. The different reasons for replacement and termination as well as the distribution of observations according to these reasons are illustrated in the tables below.

Reasons to replace	Proportion replacement for each reason
1/ vehicule change	36,8%
9/ tariff modification	16,5%
3/ reduction of guarantees	15,4%
4/ change of area	12,9%
2/ increase of guarantees	6,4%
5/ change of marital status/usage/profession	6,1%
7/ adding a clause	3,8%
6/ change of split	1,5%
8/ withdrawal of a clause	0,7%

Reasons to terminate	Proportion terminations for each reason
Sale	72,0%
Hamon	15,7%
Chatel	2,8%
Replace by	2,7%
Maturity	2,2%
Other cases	1,4%
Claim	1,1%
Due to suspension	1,0%
Change of situation	0,4%
Death	0,4%
Total loss	0,3%
Increase refusal	0,0%
Compulsory liquidation	0,0%

Breakdown of replacements and terminations following replacement by reasons

Source : AXA's portfolio (Excel)

Creation of indicators and discretization of continuous variables

Regarding the database variables, some indicators have been created to potentially increase the explanatory nature of the model. The main indicators are as follows :

— **Evolution of the technical coefficient** : the technical coefficient, called the CT, is the ratio between the premium paid and the premium for the new business. It makes it possible to identify a possible discount applied at a point in the life of the contract but also the evolution of the tariff over the years. The objective is to reduce the technical coefficient to 1 when the latter is lower than this value. Indeed, in the new business, we have

$$CT = \frac{paid\_premium}{new\_business\_premium} = \frac{new\_business\_premium}{new\_business\_premium} = 1$$

— **Contribution difference** : this variable simply indicates the difference between the contributions after and before replacement. It is expressed in euros. A negative value indicates that the contribution before is higher than that after and vice versa.

— **Replacement effective month** : this indicator can help identify replacement's seasonality. Indeed, motorcycle riding is more common during the summer months, which are synonymous with good weather, than during the winter months. Customers can therefore potentially make a replacement in early summer to increase their warranties and make a second replacement at the end of summer to decrease their warranties, with the motorcycle remaining parked in a garage.

— **Number of replacements** : This indicator counts the number of replacements made on the contract during the period from 2017 to 2019. It has been identified that some contracts had more than 10 replacements over 3 years, which is significant.

In order to obtain an explanatory coefficient per modality, a discretization of the continuous variables is necessary. This technique makes it possible to group the modalities of the continuous variables into different classes of values represented by an interval. The tool used in this study to carry out this distribution is the quantile method. Once the number of quantiles has been determined, the continuous variable is distributed according to the classes created by these quantiles. If the termination rate is roughly the same between two classes, then they can be grouped together.

Study of correlations between variables

The study of the correlation between two variables is different depending on whether they are quantitative or qualitative. The correlation between quantitative variables ([11]*RAKOTOMALALA, 2017*) is measured using the Pearson coefficient, which is 1 if the two variables are perfectly correlated, -1 if the two variables are perfectly anti-correlated, and a value between these two limits otherwise. The significance of this correlation is determined using the T-test, which indicates whether the Pearson coefficient is significantly different from 0. Within this thesis, since continuous variables have been discretized, an illustration of the correlations before discretization has been presented for illustrative purposes only. Only the correlations between qualitative variables will be retained thereafter. They are determined using the Chi-square test of independence and the intensity of this relationship is measured using Cramer's V.

Several important correlations between the variables were identified and the removal of one of them was necessary. In order to choose the best variable for the model, an XGBoost ([9] *MELLO, 2020*) was implemented on the total basis to determine the importance of each of the variables. The least important correlated variable according to the XGBoost was therefore removed for each pair and a total of 13 variables were removed. Finally, the correlation between the qualitative variables and the target variable, termination following replacement, was calculated. Those which are the least correlated with the variable to be predicted can be excluded from the study, which is notably the case for the change in splitting.

Modeling, quality measures and results

Modeling allows translating a real situation into mathematical language. In this study, the variable to be modeled is binary, i.e. it takes only two values : 1 if the contract is terminated within 180 days following the replacement, 0 otherwise. To identify the characteristics of terminations resulting from replacement and predict future ones, the supervised learning technique is used. It separates the initial database into two sub-samples : the first is the training sample, on which the model will learn to recognize terminations following replacement, and the second is the test sample, on which the model will replicate what it has learned. This separation into two sub-samples is essential so that the model can be tested under real conditions, i.e. without knowing the information in advance.

The generalized linear model

A suitable technique to model the target variable is the logit generalized linear model. This model was chosen because it allows to model binary variables and is easy to interpret, unlike probit for example. The quality of the modeling is measured by convergence criteria such as deviance, the Akaike information criterion and the Bayesian information criterion. These three criteria must be as low as possible for the model to be the best possible. Interpretation of these quality measures cannot be done alone, they must be compared to other models.

GLM logit model			
AIC	BIC	Deviance	Loglik
51 690	53 916	59 116	-25 618

Quality measures of logit GLM

Source : R (Excel layout)

The model was then applied to the test sample to predict the replacements most likely to terminate within 180 days. A termination probability was assigned to each replacement and a binarization of the replacement was required. To do this, two approaches were used. The first consists in determining the optimal prediction threshold in terms of sensitivity, proportion of correctly identified terminations, and specificity, proportion of correctly identified non-terminations. The threshold obtained is 0.06 which implies that when the probability of termination is less than 0.06 then it is transformed into 0, not terminated, and when it is greater than or equal to 0.06 then it is transformed into 1, terminated. The second approach consists in arbitrarily setting ten thresholds, from 0.05 to 0.5 in steps of 0.05.

The table below indicates for each of the thresholds the number of correctly predicted non-terminated contracts (TN), the number of poorly predicted terminated contracts (FN), the number of poorly predicted non-terminated contracts (FP) as well as the number of contracts. terminated well predicted (TP). Various performance measures were also calculated to determine which are the best prediction thresholds and which are those to be discarded. The F1-measure allows a trade-off between specificity and sensitivity : the higher the measure, the better the model. These indicators must be analyzed together. Indeed, maximizing the precision and minimizing the error rate would be the same as choosing the 0.5 threshold. Nevertheless this threshold makes it possible to predict non-terminated contracts almost perfectly and to capture only 0,4% of terminated contracts, which corresponds to those we wish to capture in this study.

Threshold	TN True negative	FN False negative	FP False positive	TP True positive	Specificity	Sensibility / Recall	Precision	Error rate	F1-measure
0,05	20 037	439	11 657	1 453	63,2%	76,8%	11,1%	36,0%	19,4%
0,06	22 010	559	9 684	1 333	69,4%	70,5%	12,1%	30,5%	20,7%
0,10	26 953	951	4 741	941	85,0%	49,7%	16,6%	16,9%	24,8%
0,15	29 499	1 309	2 195	583	93,1%	30,8%	21,0%	10,4%	25,0%
0,20	30 647	1 569	1 047	323	96,7%	17,1%	23,6%	7,8%	19,8%
0,25	31 206	1 708	488	184	98,5%	9,7%	27,4%	6,5%	14,4%
0,30	31 477	1 781	217	111	99,3%	5,9%	33,8%	5,9%	10,0%
0,35	31 592	1 828	102	64	99,7%	3,4%	38,6%	5,7%	6,2%
0,40	31 643	1 860	51	32	99,8%	1,7%	38,6%	5,7%	3,2%
0,45	31 677	1 877	17	15	99,9%	0,8%	46,9%	5,6%	1,6%
0,50	31 685	1 885	9	7	100,0%	0,4%	43,8%	5,6%	0,7%

GLM logit predictions and performance measures

Source : R (Excel layout)

In order to determine the generalization capacity of the GLM model and its stability, a k-fold cross validation was carried out. This iterative technique allows to use each of the base observations in the training sample and in the test sample. A 10-folds cross-validation was conducted which means that for each of the 10 iterations, a generalized linear logit model was implemented on different train and test samples. On average over the 10 iterations, the optimal threshold is 0.6, specificity 70,8%, sensitivity 70,5%, precision 12,8%, error rate 29,2% and F1-measurement 21,7%. These results are very close to those of the initial GLM logit which confirms that the model is stable and can be generalized.

The step procedure

A second modeling, based on GLM logit, was then conducted to challenge the results. A second modeling, based on GLM logit, was then conducted to challenge the results. The step procedure is a tool for selecting variables according to a performance criterion, in this thesis, the AIC. Step by step, the variables that increase AIC are removed from the model, so that only the most important variables are retained. There are three different stepwise methods : forward, backward and stepwise.The backward method is the one implemented here.

It allows, from the complete model that contains all the explanatory variables, to eliminate the variable which increases the AIC at each iteration. This direction was favoured because a first selection of the variables was made upstream. This direction was favoured because a first selection of the variables was made upstream. The variables taken as input by the step are therefore more likely to be explanatory of the model and few of them should be removed. Four variables of the base were removed and allowed to improve the AIC by 14 points. The BIC was also slightly improved thanks to the step procedure. Deviance remained stable and log-likelihood deteriorated.

Step model			
AIC	BIC	Deviance	Loglik
51 676	53 373	59 116	-25 665

Quality measures of step

Source : R (Excel layout)

As with the GLM logit, the step was applied to the test data to determine the probability of termination of each new replacement. Binarization was performed using the same two approaches and the optimal threshold in terms of sensitivity and specificity is also 0,06. The performance measures of the step model seem to give the same conclusions as for the GLM logit, namely that the best predictions are between 0,05 and 0,20. The quality measures for the optimal threshold are slightly better for the step than for the logit.

Threshold	TN True negative	FN False negative	FP False positive	TP True positive	Specificity	Sensibility / Recall	Precision	Error rate	F1-measure
0,05	19 969	432	11 725	1 460	63,0%	77,2%	11,1%	36,2%	19,4%
0,06	22 358	579	9 336	1 313	70,5%	69,4%	12,3%	29,5%	20,9%
0,10	26 891	967	4 803	925	84,8%	48,9%	16,1%	17,2%	24,3%
0,15	29 560	1 315	2 134	577	93,3%	30,5%	21,3%	10,3%	25,1%
0,20	30 675	1 575	1 019	317	96,8%	16,8%	23,7%	7,7%	19,6%
0,25	31 235	1 713	459	179	98,6%	9,5%	28,1%	6,5%	14,2%
0,30	31 477	1 782	217	110	99,3%	5,8%	33,6%	6,0%	9,9%
0,35	31 600	1 836	94	56	99,7%	3,0%	37,3%	5,7%	5,5%
0,40	31 646	1 859	48	33	99,8%	1,7%	40,7%	5,7%	3,3%
0,45	31 675	1 873	19	19	99,9%	1,0%	50,0%	5,6%	2,0%
0,50	31 688	1 887	6	5	100,0%	0,3%	45,5%	5,6%	0,5%

Step predictions and performance measures

Source : R (Excel layout)

Cross validation could not be implemented for the step model because of the execution time of this one. The initial step procedure took several hours to produce results. As a result, the GLM logit model is used in this study to predict termination following replacement. This choice was also motivated by the fact that the performances of the models are quite similar and that the GLM logit model has been shown to be generalizable, unlike the step.

Practical impacts of the study on retention

The study of the impact of modeling on turnover is necessary to justify its interest. Based on the confusion matrix of the table *GLM logit predictions and performance measures* presented above, a comparison of retention without tariff measure and with tariff measure has been initiated. The amounts of portfolio premium and termination premium have been modified for confidentiality reasons. The proportions have nevertheless been maintained. Before terminations, turnover amounted to 7,9M. If no pricing measures are applied, the latter will fall by 1Maccording to our predictions. In order to limit departures and thus increase retention, pricing assumptions have been applied. These are determined arbitrarily and will have to be optimized during a future study on the motorcycle perimeter. A 5% premium reduction is applied when contracts are predicted to be terminated. The retention rate of contracts that have actually terminated (*true positive*) is estimated at 75% after this benefit. Nevertheless, some terminations will not allow the insured to be retained, even with a tariff reduction, particularly in the case of the death of the driver or a compulsory liquidation. A maximum of 64% of terminations can be kept. The retention rate in this case is therefore 48%.

	TURNOVER EVOLUTION		
Thresholds	Turnover with measures / turnover without measures - 1	Turnover with measures / turnover before terminations - 1	Turnover without measures / turnover before terminations - 1
0,05	3,2%	-9,7%	-12,5%
0,06	3,1%	-9,8%	-12,5%
0,10	2,5%	-10,2%	-12,5%
0,15	1,7%	-11,0%	-12,5%
0,20	1,0%	-11,6%	-12,5%
0,25	0,6%	-12,0%	-12,5%
0,30	0,4%	-12,1%	-12,5%
0,35	0,2%	-12,3%	-12,5%
0,40	0,1%	-12,4%	-12,5%
0,45	0,1%	-12,4%	-12,5%
0,50	0,0%	-12,4%	-12,5%

Comparison of changes in turnover

Source : R (Excel layout)

Finally, the tariff measure applied allows to keep up to about 3% of the turnover. Nevertheless, these results should be analyzed with caution. Indeed, the tariff measures as well as the retention rate were determined in a totally arbitrary manner. A study of customer price sensitivity is necessary on the perimeter. Moreover, no optimum seems to have been identified in this case with the selected parameters. Nevertheless, the study has enabled us to understand the structure of the replacements and to predict those most likely to terminate within 180 days. The operational implementation of this study will be carried out after a more detailed analysis of the best pricing measure to implement.

Remerciements

Ce mémoire marque la fin de mes six années d'études supérieures dans le domaine des mathématiques appliquées. Cette dernière année que j'ai effectuée au sein du mastère spécialisé Actuariat de l'ENSAE fut un réel challenge tant sur le plan personnel que professionnel. Le contexte sanitaire que nous subissons ainsi que le confinement en mars dernier ont rendu mon départ à l'étranger incertain et la recherche d'un nouveau stage nécessaire. L'annulation de ce stage fut un coup dur face auquel j'ai réussi à rebondir grâce à mon entourage et à l'ensemble des personnes ci-après que je souhaite remercier tout particulièrement.

Je souhaite dans un premier temps remercier Alix RAMBAUD, ma tutrice au sein d'AXA France pour sa confiance et sa bienveillance tout au long de mon stage. Grâce à son expertise et son expérience, j'ai développé ma capacité à interpréter des résultats d'un point de vue opérationnel et non plus seulement mathématique. Je tiens également à remercier Andreea OLARU et Florian GAIARDO, mes collègues au sein de l'équipe moto, pour leur bonne humeur, leur disponibilité et leur patience. Plus largement je remercie l'ensemble de la direction de l'Offre IARD PP d'AXA France ainsi que RUN Services pour leur accueil.

J'aimerais ensuite remercier les différents intervenants du mastère spécialisé Actuariat de l'ENSAE, tant l'équipe pédagogique et la direction que les professionnels qui ont animé des conférences tout au long de l'année. Je souhaite particulièrement remercier Christian-Yann ROBERT, mon tuteur académique, pour ses conseils et sa réactivité, mais également Elisabeth ANDREOLETTI-CHENG pour son accompagnement tout au long de mon stage.

Je tiens également à remercier Arthur CHARPENTIER, auprès de qui je devais initialement effectuer mon stage à l'étranger, pour m'avoir mise en relation avec Alix.

Enfin, il est important pour moi de remercier mes parents, Olivier et Christine THILL, et mon binôme, Marine THUILLIER, pour leur soutien quotidien indispensable tout au long de cette année. Ils constituent un exemple de force, de détermination et de courage, qui sont les valeurs que j'ai mises à profit pour réussir cette dernière année d'études et ce stage.

Table des matières

Introduction

La motocyclette, plus communément appelée moto, est un moyen de déplacement de plus en plus utilisé par les particuliers. Initialement utilisés comme engin de déplacement utilitaire, les deux roues sont aujourd'hui principalement considérés comme des véhicules de loisir. Les risques associés à la pratique de la moto ont donc bien évolué avec le temps et les assureurs ont dû adapter leur offre. L'arrivée sur le marché de l'assurance de nouveaux acteurs, tels que les bancassureurs, a renforcé le climat concurrentiel déjà existant. L'enjeu pour les assureurs moto est donc double : augmenter leur volume de souscription tout en ayant une bonne rétention de leurs clients en portefeuille.

L'amélioration de la rétention peut intervenir à différents moments de la vie d'un contrat, notamment au terme et au remplacement. De façon plus précise, le terme correspond à la date anniversaire du contrat souvent appelée échéance. C'est à ce moment que la prime d'assurance est revalorisée. Le remplacement quant à lui correspond à une modification du contrat, en matière de garanties ou de risque associé, à n'importe quel moment, pouvant engendrer une augmentation ou une baisse de la prime d'assurance. Contrairement au terme, le remplacement est imprévisible et peut intervenir plusieurs fois dans la même année, engendrant potentiellement plusieurs augmentations de prime. L'assuré sera plus enclin à résilier son contrat en cas d'augmentation successive de sa prime au moment du terme et au moment du remplacement. L'étude menée consiste à identifier les contrats les plus susceptibles de résilier dans les six mois qui suivent le remplacement afin de mettre en place des mesures commerciales et tarifaires permettant d'améliorer la rétention de ces clients et donc de chiffre d'affaire.

L'entité motos et véhicules de collection de la Direction de l'Offre IARD Professionnels Particuliers d'AXA France pilote l'ensemble du portefeuille moto, tant au moment du terme qu'à celui des remplacements. Le terme constitue un enjeu très important du fait de sa périodicité et les études qui lui sont liées sont de ce fait prioritaires. L'étude des remplacements est donc passée en second plan et n'a jusqu'à présent pas été menée de façon précise. Ce projet va permettre de comprendre leur structure et d'identifier les différents leviers possibles pour améliorer leur pilotage.

Après avoir défini le contexte général de l'assurance et plus spécifiquement celui de l'assurance moto et de l'offre proposée par AXA en collaboration avec RUN Services, les objectifs de l'étude seront énoncés. Pour répondre à ces objectifs une base de données contenant l'ensemble des remplacements sur un historique de trois ans sera constituée. L'identification des valeurs aberrantes ainsi que la transformation des modalités *not available (NA)* permettront d'obtenir une base de données fiables et adéquates.

Ensuite, l'étude du délai entre le terme et la résiliation permettra d'identifier quelles sont les résiliations consécutives à un remplacement. C'est cet évènement qui sera modélisé et prédit par la suite. Des indicateurs tarifaires et client seront ajoutés dans la base en tant que variables potentiellement explicatives. Les variables et indicateurs continus seront ensuite discrétisés à l'aide de la méthode des quantiles pour obtenir une répartition des valeurs par classes. Les analyses univariées permettant de comprendre la structure du portefeuille de remplacement seront suivies d'une analyse précise des corrélations entre les variables ([11]*RAKOTOMALALA, 2013*). Cette étape indispensable permet de supprimer le biais engendré par deux variables trop corrélées.

La théorie des modèles linéaires généralisés, et plus précisément celle du modèle logit ([15] *ROUVIERE, 2017 et [13] RAKOTOMALALA, 2017*), sera explicitée afin de comprendre les fondements de la modélisation binaire. La procédure step ainsi que les grands principes de l'apprentissage supervisé seront ensuite présentés avant d'expliquer comment binariser les prédictions obtenues par le GLM sur l'échantillon test. Les mesures de stabilité et de qualité des modèles seront ensuite définies mathématiquement et graphiquement. La mise en pratique de ces différentes notions théoriques permettront d'évaluer quel est le meilleur modèle de prédiction des résiliations consécutives à un remplacement.

Pour finir, l'intérêt opérationnel de l'étude sera détaillé ainsi que les différentes mesures tarifaires à mettre en place pour augmenter la capacité de rétention au moment du remplacement.

Partie I

Contexte de l'étude et base de données

Le contexte général de l'assurance ainsi que les spécificités qui lui sont propres doivent être dans un premier temps abordés afin de comprendre l'environnement de l'étude. Le contexte plus particulier de l'assurance moto, objet de ce mémoire, sera ensuite défini ainsi que l'offre d'assurance proposée par AXA France sur ce périmètre. Par la suite, la construction d'une base de données fiables et exhaustives sera initiée avant de présenter les premières analyses descriptives de l'étude.

Les bases ont été extraites à l'aide du logiciel SAS. Les analyses ont été effectuées sur le logiciel R. La mise en forme des sorties a été effectuée sur Excel.

Chapitre 1

Contextes de l'étude

1.1 Contexte général

L'assurance a été créée dans le but de protéger le patrimoine des individus lors de la réalisation d'un risque. Elle permet de rendre à l'assuré le même niveau de richesse dont il disposait avant la survenance de ce risque. Les premières formes sont apparues dès l'Antiquité mais c'est à partir du XIV^{ème} siècle que l'on retrouve un système d'assurance plus proche de celui que l'on connait aujourd'hui. A cette époque, dans le commerce maritime, les marchands souscrivaient un prêt à la grosse aventure auprès des banquiers pour assurer leurs expéditions. Si le bateau faisait naufrage, les marchands ne remboursaient pas leur dette ce qui signifiait que l'expédition ne leur avait rien coûté mais également rien rapporté. A l'inverse, si la cargaison arrivait à bon port, le banquier était remboursé en totalité et bénéficiait d'une compensation financière pour faire face au risque qu'il avait pris.

Le monde de l'assurance a bien évolué depuis et celle que l'on connaît aujourd'hui est apparue suite aux tragiques incendies de Londres de 1666. C'est à cette époque que la notion de contrat d'assurance a été introduite et plus tard, la création des premières sociétés d'assurance. La protection automobile est quant à elle, apparue au début du XXe siècle. Aujourd'hui, l'assurance peut être définie comme une opération par laquelle une partie, l'assureur, s'engage, moyennant le paiement d'une prime, à verser à l'autre partie, l'assuré, une prestation lors de la réalisation d'un risque. Il existe différents domaines d'application de l'assurance notamment la vie, la non vie, la prévoyance, la mutuelle et la réassurance. Ce mémoire traitera uniquement de l'assurance non vie, aussi appelée IARD, Incendies Accidents Risques Divers. Elle correspond principalement à l'assurance des biens et des responsabilités des individus à la différence de l'assurance vie qui assure la vie de l'individu en tant que telle. En non vie, la réalisation d'un sinistre doit être extérieure et indépendante de l'assuré, aléatoire et le coût de ce sinistre n'est pas connu à l'avance. A l'inverse en assurance vie, l'élément déclencheur étant la mort de l'individu, le sinistre est certain. L'aléa intervient donc sur la date de survenance de celui-ci.

Toute la difficulté du processus de tarification en assurance réside dans le cycle inversé de production. L'assureur encaisse la prime avant même de savoir combien va lui coûter son risque, ce qui est différent d'un cycle de production classique. *Par exemple, lorsqu'un concessionnaire automobile souhaite fixer le prix d'un véhicule, il sait exactement combien lui a coûté ce dernier. En ce sens, le prix fixé tient compte du coût réel du véhicule et du bénéfice que souhaite en tirer le concessionnaire. Cela correspond à un cycle classique de production.* En assurance, il est impossible de savoir, au moment de la souscription d'un contrat, le coût que peut engendrer un assuré. *Pour revenir à l'exemple du concessionnaire, le cycle inversé de production reviendrait à fixer le prix du véhicule sans savoir combien va coûter la fabrication de celui-ci.*

Afin d'estimer au mieux le coût que peut représenter un risque pour l'assureur, un nombre important d'informations est requis au moment de la souscription, notamment les caractéristiques du bien assuré, les caractéristiques de l'individu et ses antécédents d'assurance. Néanmoins, il existe une asymétrie d'informations entre les deux parties, l'assureur et l'assuré, à différents moments de la vie du contrat qui peuvent fausser l'estimation. On parle alors d'antisélection et d'aléa moral. L'antisélection correspond au déséquilibre d'informations intervenant au moment de la conclusion du contrat. Les assurés ont tendance à surestimer la valeur de leurs biens afin d'obtenir les garanties les plus élevées possibles et donc des prestations plus importantes. L'aléa moral quant à lui intervient après la

conclusion du contrat et correspond à un changement d'attitude de l'assuré. Il aura tendance à faire moins attention au bien couvert par un contrat d'assurance que lorsque ce même bien n'était pas assuré.

C'est en ce sens qu'une mutualisation des risques est nécessaire. De façon simpliste et caricaturale, la mutualisation des risques permet d'utiliser les primes des uns pour payer les sinistres des autres. En d'autres termes, les primes collectées par les sociétés d'assurance vont servir à financer les sinistres de quelques assurés, ces derniers n'ayant heureusement pas tous des sinistres. Si la mutualisation des risques n'existait pas, l'assureur serait obligé de demander à son assuré une prime équivalente au montant qu'il pourrait potentiellement avoir à payer en cas de sinistre. La cotisation serait donc bien trop élevée et le client ne viendrait pas s'assurer. Aussi, il est important de noter que la prime d'assurance payée par l'assuré est définitivement acquise par l'assureur. Si le risque que le contrat couvre ne survient pas alors cette dernière n'est pas remboursée.

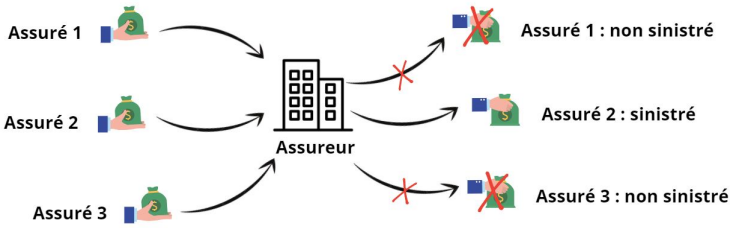


Figure 1.1.1 : Représentation de la mutualisation des risques

Source : Excel et Flaticon

Vous l'aurez compris, le calcul de la prime d'assurance est un enjeu de taille. Elle est déterminée dans les directions techniques des sociétés d'assurance et notamment dans les équipes actuarielles. Néanmoins, les actuaires ne se limitent pas au calcul de la prime. Ils analysent les contrats pendant toute la durée de leur présence en portefeuille, ils modélisent les comportements des assurés afin de prédire d'éventuels mouvements dans les contrats et ils cherchent à maintenir l'équilibre entre rentabilité et développement commercial, tant pour attirer de nouveaux clients que pour conserver ceux déjà en portefeuille. Ces différentes actions sont menées sur des segments de risques nombreux, indépendants et homogènes. Dans le cadre de ce mémoire, les analyses porteront sur le périmètre de l'assurance IARD et plus spécifiquement l'assurance automobile deux roues.

1.2 Contexte spécifique à l'étude

Les motos constituent un marché bien spécifique de l'assurance des véhicules motorisés. En auto 4 roues, la relation qu'a l'assuré avec son auto est plutôt semblable d'un individu à l'autre : dès l'obtention de son permis de conduire, l'individu assure une auto qu'il utilisera comme moyen de transport en général tout au long de sa vie. En moto, cette relation est moins systématique et évolue avec les envies et les moyens des assurés. D'un point de vue historique, les motos ont initialement été utilisées comme engin de déplacement utilitaire jusqu'à la seconde Guerre Mondiale, avant de devenir un symbole de liberté dans les années 70. Aujourd'hui, elles sont utilisées par les passionnés qui se retrouvent pour de longues balades, par les sportifs sur les circuits et les compétitions de motocross ou encore comme simple moyen de transport pour les trajets du quotidien. Les usages sont donc multiples et les risques associés différents. Cette évolution est également visible au niveau de la structure du parc de véhicules. La figure 1.1.2 représente graphiquement ce changement radical : en 1994, les cyclomoteurs représentaient près de 70% du parc de véhicules deux roues contre 20% aujourd'hui. Ils ont été majoritairement remplacés par des motos avec une cylindrée supérieure à 125cm3. Les usagers ont donc délaissé les véhicules à petite cylindrée pour gagner en puissance et donc en vitesse et en confort.

La **réglementation** a elle aussi bien évolué depuis 1994 dans le but de favoriser la sécurité routière. Afin de pouvoir conduire un deux-roues motorisé, il est nécessaire de détenir un permis qui dépend de la cylindrée du véhicule et de l'âge de son conducteur. Il existe cinq permis différents que sont :

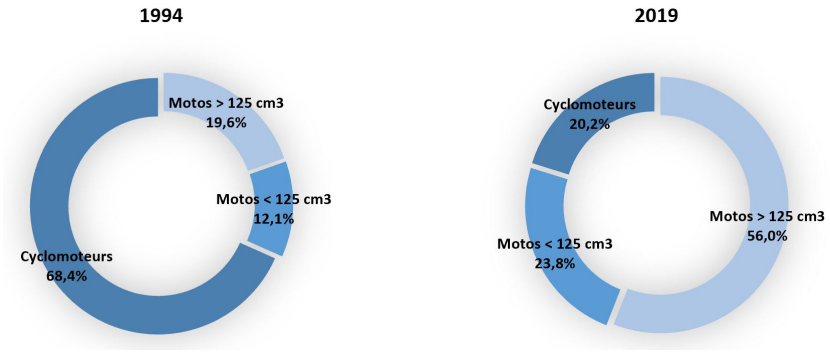


Figure 1.1.2 : Evolution du parc de véhicule de 3eme catégorie

Source : Fédération Français de l'Assurance

— **Le permis AM ou BSR** : le brevet de sécurité routière n'est pas considéré comme un permis en tant que tel mais une certification nécessaire pour conduire un cyclomoteur de 50cm3 maximum. Il est possible de passer le permis AM dès l'âge de 14 ans.

— **Le permis A1** : il permet de conduire un véhicule deux ou trois roues à moteur de 125cm3 maximum et est disponible à partir de 16 ans.

— **Le permis A2** : il est possible de passer ce permis à partir de 18 ans pour conduire des motos d'une puissance inférieure à 35kW.

— **Le permis A** : depuis janvier 2013, ce permis est accessible aux individus de plus de 24 ans qui ont 2 ans d'ancienneté de permis A2. Il permet de conduire des véhicules d'une puissance inférieure à 75kW.

— **Le permis B** : le permis auto permet de conduire des véhicules deux ou trois roues de cylindrée inférieure à 125cm3. Il est néanmoins nécessaire d'avoir 2 ans d'ancienneté de permis et d'effectuer une formation obligatoire.

Pour la pratique de la moto, il est également indispensable de disposer de l'équipement minimum nécessaire à savoir un casque homologué, obligatoire depuis 1980 pour le conducteur et le passager, ainsi que des gants certifiés par la norme européenne, depuis 2016. De nombreux autres équipements sont bien évidemment recommandés pour utiliser ce moyen de transport en toute sécurité comme par exemple les chaussures montantes ou encore le gilet airbag. En effet, à la différence de l'automobile, la moto n'a pas de carrosserie et ne protège donc ni son conducteur ni son passager. Les équipements facultatifs sont donc fortement conseillés pour une protection optimale en cas d'accident.

Aussi, d'un point de vue législatif, il est obligatoire d'assurer une moto, au même titre qu'une automobile, dans le cadre de la responsabilité civile. Cette obligation persiste même si le véhicule n'est pas utilisé et stationne dans un garage. En effet, si la moto laissée dans un garage explose, le propriétaire est responsable des dégâts occasionnés et en cas de véhicule non assuré, les conséquences financières peuvent être très importantes.

Une autre obligation légale, cette fois pour l'assureur, est de déterminer un coefficient de bonus/malus propre à chaque individu. Ce dernier est calculé annuellement en fonction du nombre d'années d'assurance et de la qualité des sinistres de l'assuré. En cas de sinistre non responsable ou lorsque l'individu n'a pas de sinistre, le coefficient de bonus/malus est diminué de 5%. Lorsque le sinistre est semi-responsable, le coefficient est majoré de 12.5%. Et lorsque l'assuré est responsable d'un sinistre, son coefficient est majoré de 25%. Lors de la première souscription d'un contrat d'assurance, le bonus/malus vaut 1. Cela signifie que l'assuré paye 100% de sa prime d'assurance. Il peut par la suite diminuer en cas de bonne conduite ou augmenter en cas de mauvais comportement du conducteur. Dans tous les cas, le coefficient bonus/malus ne peut être inférieur à 0.5 et ne peut excéder 3.5. En d'autres termes, l'assuré paye au minimum 50% de sa prime et au maximum 350% de sa prime. Ce coefficient est acquis par l'assuré et est transmis par l'assureur à la nouvelle compagnie en cas de changement de société d'assurance.

D'un point de vue **économique**, les véhicules de 3ème catégorie désignent les véhicules de moins de 4 roues à moteur donc les cyclomoteurs, motocyclettes et motocyclettes légères. Ces moyens de déplacement représentent 8,2% ([6] *Fédération Française de l'Assurance, 2019*) de l'ensemble du parc de véhicules assurés qui est également

N° client	Bonus/Malus précédent	Nb sininistres responsables	Nb sininistres non responsables	Nouveau Bonus/Malus	Prime avant application B/M	Prime après application B/M
00034543	0,50	0	0	0,50	200,00 €	100,00 €
00463745	1,00	0	1	0,95	150,00 €	142,50 €
00346793	1,10	1	1	1,38	140,00 €	192,50 €
00576890	3,50	1	0	3,50	120,00 €	420,00 €

Figure 1.1.3 : Explication du fonctionnement du bonus malus

Source : Excel

composé des véhicules de première catégorie, des flottes d'entreprises et des autres véhicules. Bien que le nombre de véhicules de 3ème catégorie augmente chaque année, la part dans le parc total de véhicules reste stable. En termes de montant de cotisations, les véhicules de 3ème catégorie ont généré en 2019 1,1 milliard d'euros soit 4,8% du total des cotisations du parc de véhicules. Afin de pouvoir comparer les données de marché avec les données AXA France disponibles au sein de la direction, seuls les véhicules de première et de 3ème catégorie ont été sélectionnés pour les statistiques suivantes. Les véhicules de 3ème catégorie représentent 9,5% de l'ensemble du parc de véhicules de première et troisième catégorie assurés sur le marché pour une part de cotisations de 5,8% du marché. Chez AXA France, les véhicules de 3ème catégorie représentent 13,4% du parc de véhicules assuré en première et troisième catégorie pour une part de cotisations de 7,6%.

Le **marché** de l'assurance moto est en forte croissance ces dernières années et devient très concurrentiel. Il fait intervenir différents acteurs tels que les compagnies d'assurance, les mutuelles et les bancassureurs qui gagnent du terrain. Dans le classement des assureurs deux roues 2019, AXA se place en 4ème position en termes de chiffre d'affaire et de nombre de contrats, derrière Generali (compagnie d'assurance), Covéa (mutuelle) et le Groupe Macif (mutuelle). Différentes stratégies sont mises en place par les acteurs du marché pour attirer un certain type de clientèle et la satisfaire. L'Assurance mutuelle des motards par exemple cible principalement les passionnés de moto alors que les acteurs plus généralistes vont plutôt cibler les détenteurs auto-moto.

1.3 Offre proposée par AXA

L'offre d'assurance moto proposée par AXA est en collaboration avec Club 14. Cette association créée en 1981 a pour objectif de rendre optimale la pratique de la moto pour ses adhérents en leur facilitant notamment l'accès à l'assurance. Dans les années 80, le manque de prévention routière pour les deux roues engendrait de nombreux accidents et les motards étaient donc des profils trop risqués pour avoir accès à l'assurance. Les primes étaient bien trop élevées et inabordables pour les usagers. C'est en ce sens qu'est né le partenariat entre le premier moto-club de France et AXA afin de proposer un tarif adapté pour le profil bien particulier que sont les motards. Deux entités complémentaires gèrent cette collaboration :

- RUN Services qui s'occupe de toute la partie gestion : gestion des contrats, gestion des sinistres, relation avec les agents, partie commerciale
- AXA qui s'occupe de la partie produits, en lien avec les équipes techniques (actuarielles), informatiques et sinistres : révision du tarif, optimisation du résultat, création de nouvelles offres en fonction des besoins énoncés par RUN. C'est au sein de la Direction de l'Offre Particuliers et Professionnels et plus particulièrement au sein de l'équipe moto, véhicules de collection, cyclo et risques aggravés, que se déroule la partie technique.

L'offre proposée par AXA et Club 14 permet d'assurer les motocyclettes légères, les motocyclettes, les tricycles à moteur et les quadricycles lourds à moteur qui peuvent être routiers ou tout terrain. Une définition plus précise des différents genres de moto ainsi qu'une représentation visuelle de ces dernières est disponible à l'annexe 3.3. Il existe six formules différentes avec des niveaux de garanties associés pour assurer un véhicule. Certaines garanties sont communes à toutes les formules et d'autres sont spécifiques à chacune.

- **Garanties communes à toutes les formules :**
  - **Responsabilité civile** (RC) : cette garantie est obligatoire et imposée par la loi. Elle permet de couvrir les dommages matériels et/ou corporels que le conducteur principal peut causer à un tiers dans le cas d'un accident impliquant le véhicule assuré.
  - **Défense Pénale et Recours Suite à Accident** (DPRSA) : cette garantie est activée dans le cas de poursuites pénales de l'assuré suite à un accident. Elle permet d'assurer la défense du sinistré.

- **Sécurité Du Conducteur de Base** (SDC base) : cette garantie permet de couvrir les préjudices corporels du conducteur en cas d'accident dont il est la victime.

— **Garanties spécifiques à chaque formule :**

- **Protection juridique** (PJ) : cette garantie fait bénéficier à l'assuré de conseils d'ordre juridique dans le domaine de la défense pénale.
- **Décès du conducteur** (DC) : cette garantie verse un capital en cas de décès du conducteur suite à un accident de la circulation.
- **Casque et gilet airbag** : ces deux garanties permettent à l'assuré de recevoir une indemnisation en cas de détérioration de son matériel lors d'un accident. Le vol du casque peut également être pris en charge dans le cadre de cette garantie sous certaines conditions spécifiques.
- **Assistance aux personnes / au véhicule** (Ass per / veh) : cette garantie prise en charge par Axa Assistance permet de venir en aide au conducteur sinistré sur le lieu de l'accident.
- **Incendie / Vol** (IV) : ces garanties sont déclenchées lorsque le véhicule a été sinistré suite à un vol ou un incendie.
- **Catastrophes Naturelles et Technologiques / Evènements climatiques / Attentat** (catnat)
- **Dommage par collision** : cette garantie s'applique lorsque le véhicule assuré a été endommagé lors d'un accident avec tiers identifié.
- **Bris d'optique** (BO) : cette garantie assure les optiques avant du véhicule.
- **Accessoires et vêtements** (Access) : cette garantie s'applique sur les éléments neufs fixés sur le véhicule par un professionnel.
- **Valeur à neuf 12 mois** (VAN) : cette garantie s'applique lorsque le véhicule assuré est détruit ou volé dans les 12 mois qui suivent sa première mise en circulation. Le terme détruit correspond au fait d'être techniquement ou économiquement irréparable.
- **Sécurité du conducteur étendue** (SDC ét.) : cette garantie correspond à la même que la sécurité du conducteur de base avec toutefois une indemnisation dix fois plus élevée.
- **Dommages tous accidents** (DTA) : cette garantie s'applique lors d'une collision entre le véhicule assuré et un ou plusieurs autres véhicules, en cas de choc entre le véhicule assuré et un objet ou dans le cas d'un acte de vandalisme.
- **Véhicule de remplacement** : cette garantie est activée à la suite d'un évènement garanti au contrat et correspond à un remboursement forfaitaire pour les frais de location de véhicule.

L'ensemble des garanties proposées par AXA et Club 14 a été énuméré. Celles-ci sont réparties dans différentes formules dont voici les principales :

— **La Tiers Essentielle** (F1) : Elle comprend les trois garanties de base (RC, DPRSA et SDC de base) ainsi que les garanties protection juridique, décès, casque et gilet airbag, assistance aux personnes et au véhicule. Il est également possible de souscrire des options telles que la PJ étendue et la SDC étendue pour être encore mieux assuré.

— **La Tiers Etendue** (F2) : Elle comprend l'ensemble des garanties de la F1 (RC, DPRSA, SDC de base, PJ (+ option PJ étendue), casque et gilet airbag, ass per/veh, SDC de base (+ option SDC étendue)) ainsi que les garanties incendie vol et catastrophes naturelles. Pour les véhicules de classe 30 [1] ou plus, les garanties bris d'optique et accessoires vêtements sont proposées en option.

— **La Tiers Sur Mesure** (F3) : Elle comprend l'ensemble des garanties de la F2 (RC, DPRSA, SDC de base, PJ (+ option PJ étendue), casque et gilet airbag, ass per/veh, SDC de base (+ option SDC étendue), IV , cat nat et pour les véhicules de classe 30 les garanties BO et Accessoires vêtements (+ option possible)) ainsi que les garanties dommages collision et valeur à neuf 12 mois pour les véhicules de classe supérieure à 30.

— **La Tous Risques Sur Mesure** (F4) : Elle correspond à la formule la plus complète et la plus souscrite par les assurés. Elle comprend l'ensemble des garanties de la F3 (RC, DPRSA, SDC de base, PJ (+ option PJ étendue), casque et gilet airbag, ass per/veh, SDC de base (+ option SDC étendue), IV , cat nat et pour les véhicules de classe 30 les garanties BO et Accessoires vêtements (+ option possible), valeur à neuf 12 mois) ainsi que les garanties dommages tous accidents, valeur à neuf 24 mois et véhicule de remplacement.

1. Les véhicules de classe 30 ou supérieure sont les véhicules routiers

L'offre d'assurance proposée par AXA et Club 14 est complète et adaptée aux besoins de chaque usager. Malgré tout, le marché de l'assurance des véhicules motorisés est très concurrentiel et le client cherche, à juste titre, à bénéficier de la meilleure protection au prix le plus avantageux. Ce climat concurrentiel est d'autant plus visible en moto lors d'une modification de contrat pouvant entraîner une hausse de cotisation et donc une résiliation. C'est en ce sens que l'étude de ces modifications de contrat fait l'objet de ce mémoire, afin de limiter la perte des clients et de chiffre d'affaire, au moment de ce fait de production.

Chapitre 2

Construction de la base de données

2.1 Objectif de l'étude

L'étude menée consiste à étudier le comportement des assurés suite à un remplacement et notamment déterminer quels sont ceux qui sont les plus susceptibles de résilier. Un remplacement correspond à un avenant, une modification du contrat. Il peut par exemple être dû à un changement de véhicule, un changement d'adresse ou un changement de catégorie socio-professionnelle. En d'autres termes, toute modification des informations du contrat nécessite d'établir un devis de remplacement. Si ce dernier est accepté par l'assuré, alors on parle de remplacement et peut engendrer une hausse ou une baisse de la cotisation annuelle d'assurance.

L'objectif pour l'assureur au moment du remplacement est de conserver son assuré tout en évitant une perte de marge. L'enjeu est donc de réajuster le niveau de tarif sur le nouveau risque tout en conservant une bonne rentabilité et une bonne rétention client. Il existe une calculette tarifaire spécifique au remplacement qui détermine la nouvelle cotisation. Elle est différente de celle utilisée au moment de l'affaire nouvelle car le client est déjà en portefeuille et n'a donc pas la même sensibilité au prix qu'un prospect. De plus, lors du remplacement, l'assuré a déjà un historique dans la compagnie qu'il faut prendre en compte lors de la retarification du contrat. De ce fait, l'ensemble des informations relatives aux remplacements et aux résiliations ont été sélectionnées mais également les informations de sinistralité et de comportement client. Les variables de sinistralité permettent d'indiquer qui est responsable du sinistre et quelles sont les garanties impactées. Les informations de comportement client indiquent par exemple le nombre de contrats détenus par l'assuré.

Les contrats d'assurance des particuliers sont annuels et renouvelés par tacite reconduction. Cela signifie que, sans intervention de l'assuré, le contrat d'assurance est renouvelé chaque année à la même date. C'est à cette date anniversaire du contrat que l'assureur revalorise les primes d'assurance, en général en les augmentant. Le calcul de cette revalorisation tarifaire à la date anniversaire du contrat s'appelle le terme. Lorsque le terme et le remplacement interviennent successivement, l'assuré peut potentiellement subir deux augmentations de tarif ce qui augmente naturellement sa propension à résilier. Le but de cette étude est donc d'identifier quelles sont les caractéristiques des remplacements qui expliquent le mieux les résiliations dans le but d'augmenter la rétention des clients tout en gardant une certaine rentabilité à terme.

2.2 Présentation de la base de données

L'une des étapes les plus importantes dans la réalisation d'une étude est la sélection des données. En effet, si les données ne sont pas en adéquation avec le périmètre à étudier, l'ensemble des analyses effectuées par la suite seront fausses et inexploitables. Il est également indispensable de vérifier que celles-ci sont pertinentes et exactes et respectent les règles de confidentialité énoncées par le RGPD [1]. Les données utilisées dans le cadre de cette étude ne permettent pas d'identifier directement les clients, les nom et prénom ne sont pas renseignés. Il est néanmoins possible de les identifier indirectement à l'aide du numéro de client ou en croisant l'ensemble des informations de la

1. Règlement Général sur la Protection des Données, CNIL

base (âge, département de résidence, situation matrimoniale...). Aucune donnée personnelle sensible n'a été utilisée pour mener à bien ces analyses.

2.2.1 Choix de l'historique

Pour analyser le comportement des clients et tenter de le prédire par la suite, il est nécessaire d'avoir accès à plusieurs années d'observations. Travailler sur une seule année peut biaiser les analyses du fait d'un manque de stabilité des données et donc de robustesse du modèle établi. Un historique de remplacements de 3 ans sur la période de 2017 à 2019 a été sélectionné pour mener à bien l'étude des remplacements résiliés. Ce choix a été motivé par le fait qu'aucune modification tarifaire majeure n'a eu lieu durant cette période, le comportement des clients n'a pas été significativement différent et les volumes d'affaires nouvelles, de remplacements et de résiliations sont similaires.

	% en fonction du nombre de contrats en portefeuille		
	Affaires nouvelles	Résiliations	Remplacements
2019	+ 16,4%	+ 17,0%	+ 15,8%
2018	+ 15,9%	+ 17,8%	+ 15,9%
2017	+ 17,2%	+ 18,7%	+ 17,3%

Figure 1.2.1 : Illustration de la stabilité dans le temps du volume de faits de production

Source : Portefeuille AXA (mise en forme Excel)

Aussi, l'historique des remplacements ne remonte pas jusqu'aux dernières données disponibles au début de l'étude, soit juin 2020, du fait du contexte économique exceptionnel lié à la pandémie de la Covid-19.

En ce qui concerne les résiliations, un historique de 3 ans et 3 mois sur la période de 2017 à la mi-mars 2020 a été sélectionné. Le 17 mars 2020 correspond à la date du premier confinement de la France des suites de la pandémie. Prendre en compte les résiliations effectuées après cette date aurait biaisé l'étude du fait de comportements atypiques des clients liés à la situation inédite de confinement.

2.2.2 Structure des bases disponibles

Les différentes bases de données utilisées sont issues du logiciel Axapac qui contient l'ensemble des informations concernant les clients et leurs contrats. L'extraction de ces bases se fait sous le logiciel SAS. Elles sont réparties en deux grandes catégories :

— Les bases portefeuille, plus souvent appelées **bases ACN**, correspondent à l'image du portefeuille à un instant t. Chaque ligne correspond à un contrat et il y figure l'ensemble des informations relatives à ce dernier à la date t choisie. Les mouvements précédents peuvent toutefois être conservés dans cette table non pas en ligne mais en colonne. Ce point sera détaillé un peu plus loin. Les bases ACN sont mises à jour régulièrement et contiennent des données fiables.

— Les bases résultat technique, plus souvent appelées **bases RT**, recensent l'intégralité des mouvements effectués sur le contrat l'année N. Chaque ligne correspond à un fait de production, mouvement effectué sur le contrat, et il est donc possible d'avoir plusieurs lignes pour un même contrat.

Le but de l'étude est de déterminer les caractéristiques des contrats résiliés suite à remplacement. Il est donc nécessaire d'avoir les informations contrat juste avant la résiliation mais également différentes informations sur le remplacement comme par exemple le motif de celui-ci ou encore l'écart entre les primes avant et après. La base finale doit comporter une ligne par remplacement avec un top 0/1 si le contrat est résilié. Dans le cas des bases RT, une ligne par fait de production, deux lignes seraient alors générées pour un même remplacement : la première indiquerait les caractéristiques du contrat avant le remplacement et la deuxième les caractéristiques du contrat après le remplacement. En choisissant les bases ACN, les informations avant et après le remplacement sont sur la même ligne.

Structure des bases ACN

N° contrat	Date effet cotis 01	Date effet cotis 02	Date effet cotis 03	Date effet cotis 04	Code évènement 01	Code évènement 02	Code évènement 03	Code évènement 04
000343045	02/12/2019	03/04/2019	02/12/2018	23/12/2017	2	1	2	1

Structure des bases RT

N° contrat	Date effet cotisation	Code évènement
000343045	02/12/2019	2
000343045	03/04/2019	1
000343045	02/12/2018	2
000343045	23/12/2017	1

Figure 1.2.2 : Structure des bases de données

Source : Axapac

Pour rentrer plus dans les détails de la construction de la base, il faut comprendre l'architecture globale d'Axapac. Les données sont réparties selon les blocs suivants :

— **Les blocs HP** : les blocs "historique de production" sont au nombre de 7. Ils recensent donc les 7 derniers mouvements effectués sur le contrat (affaire nouvelle, remplacements, suspensions, résiliation...). Le numéro 1 correspond à la situation la plus récente. C'est dans ces blocs qu'il est possible d'identifier les remplacements ainsi que leurs dates d'émission et d'effet.

— **Les blocs GC** : les blocs "garanties et cotisations" sont au nombre de 4. Ils recensent l'ensemble des informations contrat et notamment les garanties et les cotisations pour chacun d'entre eux. C'est également dans ces blocs que l'on retrouve le fractionnement, le montant du crédit commercial agent ainsi que les différentes clauses qui sont appliquées.

— **Les blocs SR** : les blocs "situation du risque" sont au nombre de 2. Ils recensent toutes les informations liées au bien assuré et à l'assuré lui-même. C'est dans ce bloc que sont par exemple renseignés le type de moto, l'immatriculation du véhicule mais aussi la date de naissance de l'assuré, sa profession et sa date de permis de conduire.

La difficulté pour récupérer les informations avant et après intervient lorsque l'individu change au moins deux fois de risque sur la même période. Dans ce cas, il est impossible d'accéder au bloc situation du risque 3, il n'existe pas. Il en est de même pour les blocs garanties et cotisations au-delà de 4 et historique de production au-delà de 7.

2.3 Construction des bases remplacements et résiliations

2.3.1 Sélection du périmètre et processus de construction

Dans un premier temps, il est primordial de sélectionner le périmètre sur lequel l'étude est effectuée. Pour ce faire, les filtres suivants ont été appliqués sur la base portefeuille ACN initiale qui recense l'ensemble des contrats des particuliers (Automobile, Multirisque Habitation,...) :

— **Faits de production et date de vision** : le premier filtre appliqué permet de sélectionner les remplacements ayant lieu sur la période de 2017 à 2019. Sur l'année 2019 par exemple, ce premier filtre permet de conserver environ 900 000 lignes.

— **Sélection du produit moto** : ce filtre permet de sélectionner le périmètre moto et réduit considérablement le nombre de lignes de la base qui devient presque 15 fois moins volumineuse.

— **Contrats temporaires et sans effet** : suppression des contrats temporaires qui correspondent à des contrats souscrits pour quelques mois uniquement dans le cas d'un véhicule peu utilisé par exemple. Le risque assuré n'est donc pas le même que le reste des contrats en portefeuille. Conserver ces contrats ne permettrait plus d'avoir une base de risques homogènes. Suppression également des affaires nouvelles sans effet et des annulations de contrat pour motif sans effet.

— **Contrats résiliés le jour de l'affaire nouvelle** : certains contrats ont une date d'effet de l'affaire nouvelle le même jour que la date de résiliation. Il est donc inutile d'inclure ces contrats dans l'étude, ils n'apportent aucune information pertinente dans ce cas.

— **Sélection des distributeurs** : le réseau d'agents généraux AXA représente la part la plus importante des distributeurs de contrats d'assurance moto. Les courtiers vendent également une partie non négligeable de

contrats. Ont été supprimés de l'étude en revanche les autres distributeurs qui sont plus spécifiques et dont les volumes de contrats en portefeuille sont très faibles.

Tous ces filtres permettent de sélectionner le périmètre de l'étude et de rendre les risques homogènes. Maintenant que le périmètre a été sélectionné, il est nécessaire de construire la base des remplacements la plus complète et la plus précise possible.

Pour ce faire, les bases portefeuille annuelles et mensuelles entre janvier 2017 et décembre 2019 ont été utilisées. Le processus de création de cette base a été le suivant :

1. **Extraction de la base annuelle des remplacements** : à partir de la base annuelle portefeuille à la vision du 31 décembre de l'année et d'un programme SAS, création d'une base contenant une ligne par remplacement avec les informations avant et après celui-ci. Il est possible d'avoir accès aux 7 derniers faits de production dans les blocs historique de production. Néanmoins, dès que l'individu a effectué 2 faits de production dans la même année, la situation du risque du deuxième fait de production ne sera pas disponible (2 blocs situation du risque uniquement). Autrement dit, lorsqu'un individu a eu plus d'un fait de production la même année, les informations ne remonteront pas dans la base annuelle, il faut aller les chercher dans les bases mensuelles.
2. **Extraction des bases mensuelles des remplacements** : grâce aux bases mensuelles, il est possible d'avoir accès à 2 blocs situation du risque par mois contre 2 blocs situation du risque sur l'année pour les bases annuelles. La récupération de l'historique des remplacements est donc envisageable, sauf pour les contrats ayant effectué 2 remplacements ou plus le même mois. L'exploitation des bases mensuelles permet également d'avoir accès aux contrats résiliés avant la fin de l'année qui n'apparaissent donc pas dans la base annuelle. Il est indispensable de prendre en compte ces contrats dans l'étude.
3. **Construction de la base finale des remplacements** : la base annuelle est utilisée comme référence initiale. Les informations manquantes des remplacements de la base annuelle sont complétées par celles disponibles dans les bases mensuelles. Ensuite, l'ensemble des contrats présents dans les bases mensuelles mais pas dans les bases annuelles, c'est à dire les contrats résiliés en cours d'année, sont ajoutés à la base de référence. La base finale des remplacements contient donc l'ensemble des contrats des bases annuelles avec les informations à jour ainsi que l'ensemble des contrats résiliés en cours d'année ayant effectué un remplacement.

Ce processus de construction de la base de données finale des remplacements permet de faire face au problème d'historisation des données et permet également de récupérer l'ensemble des remplacements effectués au cours de l'année.

Afin d'identifier les contrats ayant effectué un remplacement puis résilié, la base des résiliations de janvier 2017 à mars 2020 a été extraite suivant le même processus que la base des remplacements. Les deux bases ont ensuite été combinées par le biais d'une clé unique. Lorsqu'un individu a effectué 2 remplacements puis résilié, seul le dernier remplacement en date sera identifié comme résilié. Parmi les 1 500 variables de cette base, beaucoup ne sont pas relatives au périmètre moto mais à d'autres périmètres comme par exemple l'assurance multirisque habitation. De ce fait, les variables n'ayant aucun rapport avec l'étude ont été supprimées, permettant de réduire le nombre de colonnes de la base à environ 350.

2.3.2 Les types de remplacements et de résiliations

Remplacements

Les motifs de remplacement ont été déterminés suite à la comparaison des caractéristiques des contrats avant et après. Ils sont hiérarchisés selon l'importance du changement effectué pour AXA.

Un assuré qui déménage et change de véhicule n'apparaîtra qu'une seule fois dans le tableau de la figure 1.2.3 sur la ligne *1/ changement de véhicule*. Un individu qui déménage suite à une séparation sera comptabilisé sur la ligne *4/ changement de zone-cdp*. Autrement dit, les motifs de remplacements sont affectés selon l'ordre des numéros par lesquels leur modalité commence. Plus d'un tiers des remplacements sont dus à un changement de véhicule.

Motifs de remplacement	Proportion de remplacements pour chaque motif
1/ changement de véhicule	36,8%
9/ modification tarifaire	16,5%
3/ diminution de garanties	15,3%
4/ changement de zone-cdp	12,9%
2/ extension de garanties	6,4%
5/ changement de sitma/usage/csp	6,1%
7/ ajout d'une clause	3,8%
6/ changement de fractionnement	1,5%
8/ retrait d'une clause	0,7%

Figure 1.2.3 : Répartition des remplacements par motifs

Source : Excel

Les modifications tarifaires peuvent être par exemple dûs à une refonte du zonier ou un changement de classification des prix de véhicules. Dans ce cas, si l'individu change uniquement son nom de famille, un mouvement de prime sera tout de même effectué. Le motif de remplacement numéro 5/ regroupe les changements de situation matrimoniale, d'usage et de catégorie socio-professionnelle.

Résiliations

En assurance, les résiliations peuvent être à l'initiative de l'assureur ou de l'assuré. L'assureur peut résilier le contrat de son assuré principalement en cas de non paiement de la prime. L'assuré quant à lui peut résilier son contrat quand il le souhaite, avec néanmoins un préavis. L'étude menée ici consiste à étudier le comportement des assurés qui résilient suite à un remplacement et donc conserver uniquement les résiliations initiées par les assurés. Les motifs présents dans la base sont les suivants ([8] *INDEX ASSURANCE, 2020*) :

- **Changement de situation** : lorsque l'assuré change de situation (matrimoniale, déménagement, catégorie socio-professionnelle...) la résiliation prend effet 1 mois après la notification et la justification du changement par l'assuré.
- **Décès** : lors du décès du conducteur principal, ce sont les ayants droits qui héritent du véhicule et donc de l'assurance. Ces derniers peuvent demander la résiliation du contrat sans préavis.
- **Échéance** : la résiliation à l'échéance permet à l'assuré d'annuler la tacite reconduction et donc d'arrêter son contrat à sa date d'anniversaire. Pour cela, un courrier recommandé avec accusé de réception au plus tard deux mois avant la date d'échéance doit être transmis à l'assureur.
- **Loi Châtel** : les résiliations loi Châtel interviennent lorsque l'assureur a oublié de notifier à son assuré qu'il pouvait résilier à la date d'échéance. Ce dernier dispose donc d'un délai supplémentaire pour arrêter son contrat.
- **Loi Hamon** : la loi Hamon permet à l'assuré, après un an d'ancienneté d'assurance, de résilier son contrat à tout moment et sans motif. Le client est simplement tenu de souscrire une couverture en responsabilité civile chez un autre assureur avant de résilier.
- **Liquidation judiciaire** : lorsque l'assuré subit une liquidation judiciaire, il peut choisir de résilier son contrat d'assurance à tout moment.
- **Perte totale** : ce motif de résiliation intervient lorsque le conducteur principal subit une perte totale de son véhicule suite à un évènement non prévu au contrat. Dans ce cas la résiliation est immédiate et sans préavis.
- **Refus de majoration** : lorsque la prime d'assurance augmente sans changement de risque, l'assuré peut résilier dans les 15 jours qui suivent la notification de la majoration.
- **Remplace par** : ce motif n'est en réalité pas une vraie résiliation. Il correspond à une résiliation suivie d'une affaire nouvelle. Cette technique est utilisée à la place d'un remplacement.
- **Sinistre** : suite à un sinistre, la prime d'assurance est recalculée et majorée. L'assuré peut décider de résilier suite à une majoration sinistre.
- **Suite suspension** : lors de la suspension de permis, les assurés peuvent être contraints de se séparer de leur véhicule et donc de résilier leur contrat d'assurance dans l'immédiat.
- **Vente** : la résiliation en cas de vente de véhicule est effective dans les jours qui suivent la vente du véhicule. Une copie du certificat de cession doit être transmise à l'assureur.
- **Autres cas**

Les deux motifs de résiliation les plus importants pour les contrats qui ont effectué au moins un remplacement sont la vente de véhicule et la loi Hamon. Ces deux motifs représentent près de 85% des résiliations suite à remplacement. La vente de véhicule représente à elle seule près de deux tiers des résiliations dans ce contexte. Les résiliations loi Châtel, à l'échéance et suite à suspension représentent presque la même proportion de résiliations. Les autres motifs sont assez peu représentés et donc plutôt rares.

Motifs de résiliation	Proportion de résiliations pour chaque motif
Vente	63,5%
Hamon	20,5%
Loi Chatel	4,0%
Echeance	3,5%
Suite suspension	3,4%
Remplace par	2,3%
Sinistre	1,0%
Autres cas	0,8%
Changement de situation	0,3%
Décès	0,3%
Perte Totale	0,3%
Liquidation judiciaire	0,0%
Refus de majoration	0,0%

Figure 1.2.4 : Répartition des résiliations par motifs

Source : Excel

2.4 Caractérisation des variables

2.4.1 Les différentes catégories de variables

Les variables présentes dans la base peuvent être regroupées en quatre grandes catégories :

— **Caractéristiques du bien assuré** : elles correspondent à l'ensemble des informations relatives à la moto telles que la cylindrée, la date de mise en circulation du véhicule ou encore le genre du véhicule. Ces variables peuvent être identifiées par le préfixe *VEH*, diminutif de véhicule.

— **Caractéristiques du client** : elles correspondent à l'ensemble des informations permettant d'identifier le client telles que la date de naissance, la catégorie socio-professionnelle ou encore la date d'obtention du permis de conduire. Ces variables peuvent être identifiées par le préfixe *PER*, diminutif de personne.

— **Caractéristiques du contrat** : elles correspondent à l'ensemble des informations relatives au contrat telles que les garanties souscrites, le fractionnement ou encore le montant de la cotisation. Ces variables peuvent être identifiées par le préfixe *POL*, diminutif de police d'assurance.

— **Comportement du client** : elles correspondent à l'ensemble des informations relatives au risque que représente le client telles que la sinistralité, l'indice de qualité ou encore les antécédents. Ces variables peuvent être identifiées par le préfixe *CLA*, diminutif de comportement lié à l'assuré.

Parmi ces grandes catégories de variables, il existe des variables dites tarifaires, c'est à dire qui jouent un rôle dans le calcul de la prime d'assurance. L'ancienneté de permis de conduire, le fractionnement, la sinistralité antérieure et le genre du véhicule font partie des variables tarifaires.

2.4.2 Les différents types de variables

Non seulement les variables peuvent être regroupées en catégorie selon l'information qu'elles apportent mais elles sont également différenciées par leur type : qualitative, quantitative ou temporelle.

Les **variables quantitatives** s'expriment par un nombre et peuvent être discrètes ou continues. Les variables quantitatives discrètes prennent uniquement des valeurs entières. La variable nombre de contrat par exemple est une variable quantitative discrète : il n'est pas possible de détenir 2,5 contrats. Les variables quantitatives continues quant à elles peuvent prendre n'importe quelle valeur entière ou décimale. La variable indiquant le montant de cotisation est par exemple quantitative continue.

Les **variables qualitatives** s'expriment par une caractéristique, un mot ou une lettre et peuvent être nominales, ordinales ou binaires. Les variables qualitatives nominales ne peuvent être hiérarchisées. La situation matrimoniale du client est une variable qualitative nominale. En effet, les valeurs prises sont des mots (marié, veuf, divorcé) et aucune de ces valeurs n'est supérieure à une autre. Les variables qualitatives ordinales peuvent, elles, être hiérarchisées. La classe de prix du véhicule est exprimée par une lettre entre A et R. La lettre A représente la classe de prix la moins élevée et la lettre R la plus élevée. Enfin, les variables qualitatives binaires prennent uniquement deux valeurs. La variable changement de véhicule prend les valeurs "O", pour oui, ou "N", pour non.

Enfin, les **variables temporelles** représentent une unité de temps comme par exemple une date. Les variables date d'effet du remplacement et date d'obtention du permis de conduire sont deux variables temporelles.

2.5 Fiabilisation des données

Un long travail rigoureux de fiabilisation des données a été nécessaire pour mener à bien une étude. Il est important d'identifier les valeurs manquantes ainsi que les valeurs aberrantes et de proposer des transformations de ces variables.

2.5.1 Identification des valeurs aberrantes

Afin de détecter la présence de valeurs aberrantes, quelques vérifications élémentaires sont nécessaires. La première vérification effectuée concerne la cohérence des données comme par exemple l'âge du conducteur principal. L'âge minimum pour conduire une moto ou un scooter est de 14 ans. Par conséquent, aucun contrat ne doit donc avoir un âge inférieur à 14. Néanmoins, une dizaine d'entre eux ne respectaient pas cette condition et avaient un conducteur principal âgé de moins de 14 ans. Le volume de ces contrats étant très faible et l'âge renseigné étant sûrement dû à une erreur d'alimentation de la base, ils ont été supprimés.
Une autre vérification élémentaire concerne les modalités NA, *not available*. A l'aide de la fonction *summary* du logiciel R, ces valeurs manquantes sont facilement identifiables. La base de données contient un grand nombre de valeurs manquantes qu'il faut supprimer ou transformer. Une analyse détaillée du traitement de ces modalités sera effectuée dans la sous-section suivante.

Aussi, les montants de cotisation et de crédit commercial agent ont été étudiés de façon plus précise. Le choix de ces deux variables en particulier a été motivé par le fait qu'elles ont une grande importance dans la tarification. La cotisation correspond à la prime payée par le client et le crédit commercial agent correspond à la réduction, en montant, accordée par l'agent général au client. Cet avantage correspond à la part en euros de la cotisation prise en charge par l'agent. Les boxplot de la figure 1.2.5 permettent d'identifier en un coup d'oeil les valeurs aberrantes, qui sont ici très élevées.

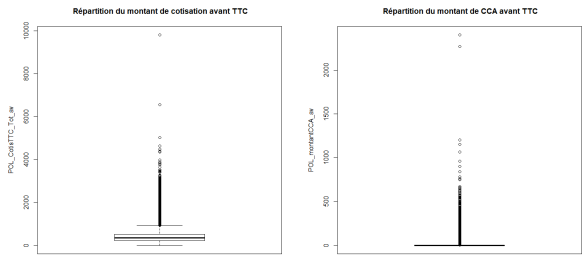


Figure 1.2.5 : Identification des valeurs aberrantes
Source : R (Boxplot)

Une étude détaillée de ces quatre valeurs a été menée notamment à l'aide du logiciel Axapac. L'investigation menée a permis de détecter que les anomalies de prime et de crédit commercial agent interviennent dans la majorité des cas lorsque l'individu a effectué un remplacement le même jour que la date anniversaire de son contrat. Dans ce cas, le logiciel n'arrive pas à historiser correctement les informations. La vision qui est censée être la plus récente car en première position dans le logiciel s'avère en réalité être la deuxième vision la plus récente. La résolution de ce problème fut assez complexe : certains contrats ont réellement vu leur prime être multipliée par 8 lors d'un remplacement. Pour d'autres, cette valeur sera aberrante. Il a donc été nécessaire d'identifier une partie de la prime fixe pour tous les contrats. La prime assistance aux personnes ne prend que trois valeurs, peu importe les caractéristiques du contrat. Lorsque le montant de cette prime différait d'une des trois valeurs qu'elle est sensée prendre, la cotisation totale, le crédit commercial agent ainsi que le fractionnement étaient modifiés dans les mêmes proportions. Ce processus d'identification et de modification des valeurs aberrantes en termes de primes, montant de CCA et fractionnement a été appliqué sur les primes avant et après remplacement.
Enfin, pour 8% des remplacements, les informations avant n'étaient toujours pas disponibles, ni dans la base annuelle, ni dans les bases mensuelles. Néanmoins, les lignes concernées n'ont pas été supprimées car elles représentaient une proportion trop importante des données. En cas de suppression, la perte des informations après remplacement aurait été trop significative.

2.5.2 Transformation des NA

Afin d'avoir une base de données exploitable, il est nécessaire de supprimer ou transformer les valeurs NA, *not available*. La première solution qui est de supprimer ces valeurs ne peut être retenue car elle ferait perdre beaucoup

trop d'informations. La transformation de cette modalité, différenciée selon le type des variables, est donc nécessaire ([16] TREMBLAY, 2017). Pour les variables quantitatives, la modalité NA est remplacée par une valeur numérique qui doit être facilement identifiable. Dans le cadre de cette étude, la valeur numérique choisie est -100. En ce qui concerne les variables qualitatives, la modalité NA a été remplacée par "NR" qui signifie non renseigné. Enfin, pour les variables temporelles, la modalité inconnue a été remplacée par une date très éloignée de celles que l'on peut trouver en portefeuille soit "1800-01-01". Le code R ayant permis d'effectuer l'ensemble de ces transformations est disponible à l'annexe 1.3.3.

Partie II

Analyses descriptives et sélection de variables

La fiabilisation des données ayant été effectuée, il est nécessaire de définir et d'identifier les résiliations suite à remplacement. Les modalités de certaines variables seront ensuite regroupées par classes afin de comprendre la structure globale des remplacements. De nouveaux indicateurs seront également créés et analysés dans le but de potentiellement mieux expliquer la résiliation suite à modification de contrat. Enfin, afin d'avoir un modèle non biaisé, les corrélations entre les variables puis avec la variable cible seront étudiées. Cette dernière étape permettra de supprimer un nombre important de colonnes dans la base.

Les modifications de variables ont été effectuées sur R. Les analyses univariées ont été effectuées sur R et mises en forme sur Excel. De même pour les corrélations.

Chapitre 1

Mise en forme des données

1.1 Étude du délai de résiliation suite à remplacement

La base de données constituée contient l'ensemble des remplacements effectués sur la période de 2017 à 2019 ainsi que les informations de résiliation qui lui sont associées, si le contrat a été résilié.

Il est nécessaire de définir un horizon de temps entre le remplacement et la résiliation afin de réaliser des prédictions intéressantes. Pour ce faire, le délai ne doit être ni trop faible ni trop important. En effet, si le délai est trop faible, peu de résiliations seront prises en compte et la capacité de prédiction du modèle sera minime. A l'inverse, si le délai est trop important, les remplacements auront plus de chance d'être résilié mais une mise en place opérationnelle nécessitera beaucoup de ressources. Afin de déterminer le délai adéquat entre le remplacement et la résiliation, ce dernier a été représenté graphiquement sur la figure 2.1.1.
Grâce à cette représentation graphique, il apparaît très clairement que pour un délai supérieur à 1 an, la résiliation ne semble pas être initiée par le remplacement. Le délai entre le remplacement et la résiliation atteint son plus haut niveau quelques semaines après le remplacement jusqu'à 180 jours après, axe représenté en pointillés rouges.

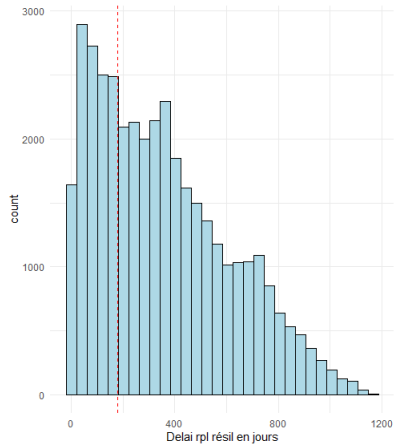


Figure 2.1.1 : Représentation du délai entre le remplacement et la résiliation en jours
Source : Logiciel R (GGplot)

Le délai choisi graphiquement est donc de 180 jours ce qui signifie qu'une résiliation est consécutive à un remplacement lorsque celle-ci intervient dans les 180 jours qui suivent ce remplacement. Afin de confirmer cette intuition graphique, une représentation chiffrée en termes de volume et de perte de chiffre d'affaire a été initiée. Différents délais ont été comparés à savoir 60, 120, 180, 240 et 365 jours.

La figure 2.1.2 ci-dessous permet d'identifier quel est le délai le plus intéressant à étudier. Il est clair que le délai de 60 jours ne présente pas assez d'enjeu économique car il n'engendre que 2,5% de la perte totale de chiffre d'affaire sur le périmètre. De même pour le délai de 120 jours qui n'engendre que 4,6% de la perte totale de chiffre d'affaire. Le délai de 365 jours est, à dire d'expert, trop important pour pouvoir prédire les résiliations. Les délais de 180 jours et 240 jours sont donc les deux candidats restants pour définir les résiliations consécutives à un remplacement. En pratique et dans les études menées au sein de la Direction sur d'autres périmètres, il est d'usage d'utiliser un délai de 3 mois ou 6 mois. Le choix final se porte donc naturellement sur un délai de 6 mois, soit 180 jours.

	Résiliations suite à remplacement à l'initiative de l'assuré entre 2017 et mars 2020 selon le délai sélectionné		Proportion par rapport au total des résiliations à l'initiative de l'assuré entre 2017 et mars 2020	
Délai entre le remplacement et la résiliation	Volume	Perte de chiffre d'affaire	Volume	Chiffre d'affaire
60 jours	3 816	1 528 134 €	2,3%	2,5%
120 jours	7 024	2 845 844 €	4,2%	4,6%
180 jours	9 922	4 013 726 €	5,9%	6,6%
240 jours	12 414	5 052 182 €	7,4%	8,3%
365 jours	17 568	7 120 502 €	10,4%	11,6%

Figure 2.1.2 : Représentation du chiffre d'affaire selon le délai entre le remplacement et la résiliation

Source : Portefeuille AXA

L'ensemble de ces éléments graphiques et numériques permettent de définir les résiliations consécutives à un remplacement comme des résiliations survenues dans les 180 jours qui suivent ce remplacement.

Afin de garder le caractère homogène de la base de données, il est nécessaire que chaque remplacement dispose de la même fenêtre d'observation, à savoir 180 jours. En effet, pour un remplacement effectué le 31 décembre 2019 et non résilié avant la mi-mars 2020, la fenêtre d'observation sera d'environ 75 jours. Si ce même contrat est résilié début avril, il devra être considéré comme résilié suite à remplacement. Néanmoins, cette information n'étant pas disponible dans la base, il sera considéré à tort comme non résilié. Ces remplacements doivent donc être retirés de la base de données et sont au nombre de 15 227. Les remplacements qui n'ont pas 180 jours de fenêtre d'observation mais qui ont résilié avant la mi-mars peuvent néanmoins être pris en compte dans l'étude. La répartition, en termes de proportions, des résiliations suite à remplacement est illustrée dans le tableau de la figure 2.1.3 ci-contre.

Motifs de résiliation	Proportion de résiliations suite à remplacement pour chaque motif
Vente	72,0%
Hamon	15,7%
Loi Chatel	2,8%
Remplace par	2,7%
Echeance	2,2%
Autres cas	1,4%
Sinistre	1,1%
Suite suspension	1,0%
Changement de situation	0,4%
Décès	0,4%
Perte Totale	0,3%
Refus de majoration	0,0%
Liquidation judiciaire	0,0%

Figure 2.1.3 : Répartition des résiliations suite à remplacement par motifs

Source : Portefeuille AXA (mise en forme Excel)

1.2 Création d'indicateurs

Afin d'expliquer au mieux les résiliations suite à remplacements, différents indicateurs ont été créés. Ils permettent notamment de rendre compte de l'évolution de certaines variables tarifaires au moment du remplacement ou caractérisent de manière plus précise les individus.

1.2.1 Indicateurs tarifaires

Les indicateurs suivants ont été créés dans le but de comparer les informations avant et après le remplacement en un coup d'oeil et d'identifier plus facilement certaines caractéristiques des contrats.

— **Évolution de la prime au remplacement** : cet indicateur permet de visualiser l'écart entre les primes avant et après remplacement en termes de proportion. Il est calculé de la façon suivante :

$$evol\_prime = \frac{Cotis\_TTC\_totale\_ap}{Cotis\_TTC\_totale\_av} - 1$$

S'il est positif, alors la prime après est plus élevée que la prime avant remplacement. A l'inverse, en cas d'indicateur négatif, la prime avant est plus élevée que la prime après.

— **Évolution du coefficient technique au remplacement** : le coefficient technique, appelé CT, correspond au ratio entre la prime payée et la prime à l'affaire nouvelle. C'est un indicateur de sur ou sous tarification du contrat. En effet, à l'affaire nouvelle, on a

$$CT = \frac{Prime\_paye}{Prime\_affaire\_nouvelle} = \frac{Prime\_affaire\_nouvelle}{Prime\_affaire\_nouvelle} = 1$$

L'évolution du coefficient technique permet d'identifier un éventuel rabais appliqué lors du remplacement.

— **Mouvement de prime** : cet indicateur permet d'identifier si le remplacement a généré une hausse de prime, une baisse de prime ou n'a pas généré de changement de prime. Une simple comparaison des primes totales TTC avant et après est effectuée.

— **Écart de cotisation** : cette variable indique simplement l'écart entre la cotisation avant remplacement et la cotisation après remplacement en euros.

— **Taux de crédit commercial agent** : grâce à cet indicateur, il est possible de déterminer la proportion de budget octroyé par l'agent au moment du remplacement. Cette variable est calculée de la manière suivante :

$$taux\_CCA = \frac{Montant\_CCA\_TTC}{Cotis\_TTC\_totale}$$

— **Délai entre le terme et le remplacement** : cette variable a été créée dans le but d'identifier un potentiel lien entre un remplacement et un terme successif qui engendrerait une résiliation. Le terme intervenant chaque année, le délai a été centré en 0. Une valeur négative indique que le terme a eu lieu avant le remplacement et une valeur positive indique que le terme a eu lieu après le remplacement.

— **Résiliation selon le mois** : cet indicateur permet de déterminer si la résiliation est intervenue dans le premier mois suivant le remplacement, le deuxième, le troisième, le quatrième, le cinquième ou le sixième.

— **Multi-détenteur** : cet indicateur permet de différencier les clients qui ont un unique contrat chez AXA, mono-détenteurs, de ceux qui en ont plusieurs, multi-détenteurs.

— **Ancienneté de contrat au remplacement** : cette variable indique le nombre de jours qui séparent la date d'effet de l'affaire nouvelle et la date d'effet du remplacement.

— **Mois d'effet du remplacement** : cet indicateur peut permettre d'identifier une saisonnalité des remplacements. En effet, la pratique de la moto est plus courante pendant les mois d'été qui sont synonymes de beau temps, que pendant les mois d'hiver. Les clients peuvent donc potentiellement effectuer un remplacement au début de l'été pour augmenter leurs garanties et effectuer un deuxième remplacement à la fin de l'été pour diminuer leurs garanties, la moto restant stationnée dans un garage.

— **Nombre de remplacements** : cet indicateur compte le nombre de remplacements effectués sur le contrat pendant la période de 2017 à 2019. Il a été identifié que certains contrats avaient plus de 10 remplacements sur 3 ans ce qui est conséquent.

1.2.2 Indicateurs sur la vision du client

Deux indicateurs supplémentaires ont été créés afin de mettre en lumière certaines incompréhensions tarifaires du point de vue du client. Par exemple lorsqu'un client change de situation matrimoniale, il peut être difficile pour lui de comprendre que sa prime d'assurance augmente. De même pour un déménagement ou une baisse de garanties qui peuvent engendrer une hausse de la cotisation. Par exemple, prenons un client ayant souscrit une affaire nouvelle en 2002 et n'ayant fait aucun mouvement sur son contrat depuis. Lorsque l'individu initie un remplacement en 2018 sur son contrat, le tarif proposé lors du devis de remplacement prendra en compte les coefficients tarifaires de 2018, et non ceux en vigueur en 2002, ce qui modifiera son tarif. Si par exemple une refonte du zonier moto a eu lieu entre temps et que la zone d'habitation du client est passée d'une zone 10 à une zone 12, l'augmentation tarifaire du fait de ce changement sera prise en compte au moment du remplacement. En outre, il se peut donc que par des effets croisés, un client voit son tarif augmenter au moment d'un remplacement alors qu'il a baissé le niveau de ses garanties. Une telle situation, bien que rare, reste à juste titre difficile à comprendre pour un assuré et peut engendrer une résiliation.

Afin d'identifier de telles situations, deux indicateurs de hausse et de baisse de prime difficilement intelligibles pour le client ont été créés. L'indicateur à la hausse prend la valeur 1 lorsqu'une hausse de prime est combinée avec un changement de situation matrimoniale, une diminution de garanties, un changement de zone ou un changement de catégorie socio-professionnelle. L'indicateur à la baisse prend quant à lui la valeur 1 lorsqu'une baisse de prime est combinée avec un changement de situation matrimoniale, une extension de garanties, un changement de zone ou un changement de catégorie socio-professionnelle.

1.3 Répartition des modalités par classes

1.3.1 Variables tarifaires

La répartition des modalités par classe pour les variables tarifaires a été effectuée sur la segmentation existante et utilisée dans le modèle tarifaire. En effet, afin de pouvoir comparer facilement les volumes obtenus et réutiliser l'étude en pratique, il est nécessaire de disposer des mêmes classes de risques sur tous les outils. Les variables tarifaires concernées sont :

- **Ancienneté de permis** : exprimée en années, elle prend ses valeurs entre 0 et 70 et correspond à la différence entre la date d'obtention du permis de conduire et la date d'effet du remplacement.
- **Age du véhicule** : exprimée en années, cette variable indique l'ancienneté du véhicule qui correspond à la différence entre la date de la première mise en circulation et la date d'effet du remplacement.
- **Durée de détention tarifaire** : exprimée en année, elle correspond à la différence entre la date d'affaire nouvelle du contrat moto et la date d'effet du remplacement. Elle indique depuis combien d'années le contrat est présent en portefeuille.
- **Durée de détention du véhicule** : exprimée en années, cette variable correspond à la différence entre la date d'établissement de la carte grise et la date d'effet du remplacement.
- **Cylindrée du véhicule** : la cylindrée du véhicule indique sa puissance et est exprimée en cm^3. Elle est différente selon le groupe de moto, tout terrain ou routier, et selon le genre, scooter ou moto. La répartition des modalités de cette variable a donc été effectuée conjointement avec le groupe et le genre.
- **Type de roues** : cette variable est renseignée uniquement pour les scooters deux et trois roues et indique si les roues sont petites ou grandes. Cette information permet de juger de l'équilibre du véhicule.
- **Age du conducteur principal** : exprimée en années, cette variable correspond à la différence entre la date de naissance du conducteur principal et la date d'effet du remplacement.

1.3.2 Discrétisation des variables continues

Afin de pouvoir appliquer un modèle en ayant un coefficient explicatif par modalité, il est nécessaire de regrouper les valeurs prises par les variables continues sous forme de classes. On parle alors de discrétisation car les variables continues vont être transformées en variables discrètes. Il existe plusieurs méthodes pour effectuer ce regroupement notamment la méthode des quantiles qui a été utilisée ici. Elle permet de répartir les observations dans des classes contenant le même nombre de données. De façon plus mathématique, le quantile d'odre P est déterminé par $q = F^{-1}(p)$ avec F une fonction de répartition et $p \in]0;1[$. Par exemple, si l'échantillon doit être divisé en déciles, alors $p = \frac{1}{10}$ et q vérifiera $F(q) = p \iff P(X \leq q) = \frac{1}{10}$. Cela signifie que le premier quantile est la valeur pour laquelle un dixième des observations lui seront inférieures. Le processus de dicrétisation par quantile des variables continues est le suivant :

1. **Sélectionner la variable à discrétiser** : la variable à discrétiser doit être continue comme par exemple la cotisation totale.
2. **Choisir le nombre de quantiles** : il est nécessaire de fixer le nombre de quantiles souhaités. Si dix quantiles sont sélectionnés, on parle alors de discrétisation par déciles. Les cotisations par exemple seront regroupées dans dix classes différentes.
3. **Regroupement de certaines classes** : afin d'avoir un nombre limité de nouvelles modalités, la probabilité de résiliation suite à remplacement a été calculée pour chacune des classes. Celles avec une probabilité proche ont été regroupées en une seule et même classe.

Les variables continues qui ont été discrétisées selon ce processus sont les suivantes :

- **Évolution du coefficient technique au remplacement** : cet indicateur tarifaire créé dans la section 1.2.1 a été réparti en 10 classes : 4 modalités représentent les évolutions négatives, 1 modalité représente une évolution nulle, 4 modalités représentent les évolutions positives et une modalité représente les valeurs manquantes.
- **Évolution de la prime au remplacement** : cet indicateur tarifaire également créé dans la section 1.2.1 a été réparti en 8 classes : 4 modalités représentent une évolution négative, 1 modalité représente une évolution nulle et 3 modalités représentent une évolution positive.

— **Taux de crédit commercial agent** : cet indicateur tarifaire également créé dans la section 1.2.1 a été réparti en 5 classes : 1 modalité recense les remplacements n'ayant pas bénéficié du budget des agents et 4 modalités représentent une évolution positive.

— **Ancienneté du contrat au remplacement** : cet indicateur tarifaire également créé dans la section 1.2.1 a été réparti en 6 classes : 1 modalité indique que le remplacement a eu lieu 1 jour après la création du contrat et 5 modalités représentent un intervalle de temps exprimé en jours.

— **Montant de crédit commercial avant** : le montant de crédit commercial a été réparti en 5 modalités : une modalité indiquant qu'il n'y a pas eu de budget octroyé par l'agent au moment du remplacement et 4 modalités qui représentent des intervalles de montants.

— **Cotisation totale** : la cotisation totale a été répartie en 4 classes.

— **Ancienneté du client** : l'ancienneté du client a été répartie en 8 modalités qui représentent des intervalles exprimés en années. Une modalité représente les valeurs manquantes.

— ***Estimated Loss Ratio*** : l'ELR correspond à une projection du ratio S/C, sinistres sur cotisations. Les sinistres sont calculés de manière prospective car inconnus au moment de la souscription. Ce ratio indique la part de la cotisation consommée par les sinistres futurs. Si l'*estimated loss ratio* est plus grand que 1, cela signifie que le client va coûter à AXA plus cher que la prime qu'il aura payée.

— **Écart de cotisation** : cet indicateur tarifaire créé dans la section 1.2.1 a été réparti en 9 modalités : 4 modalités négatives qui indiquent que la prime avant est plus élevée que la prime après remplacement, 4 modalités positives qui indiquent que la prime avant est moins élevée que la prime après remplacement et une modalité nulle qui indique que la prime n'a pas changé au moment du remplacement.

Chapitre 2

Analyses univariées

Les analyses univariées permettent de décrire la structure des données et notamment la répartition des individus dans chacune des modalités. Une représentation chiffrée, à l'aide d'un tableau de données, permet d'avoir accès au volume exact de chaque modalité de la variable tandis qu'une représentation graphique permet d'identifier d'un coup d'oeil les modalités qui résilient le plus suite à remplacement. Ces deux représentations complémentaires permettent d'avoir accès à l'ensemble des informations disponibles.

Dans le cadre de ce mémoire, les analyses univariées ont été menées sur les données avant et après remplacement pour identifier les mouvements d'une modalité à l'autre lors de ce fait de production et donc une éventuelle modification de structure. Aucune modification majeure de structure au moment du remplacement n'a été décelée et seules les analyses après remplacement seront présentées ici.

Le taux de résiliation a été représenté graphiquement à l'aide d'une courbe et d'un intervalle de confiance à 95%. L'intervalle de confiance est calculé à l'aide des formules suivantes

$$IC_{5\%} = p - z * \sqrt{\frac{\sigma^2}{n}} = p - 1,96 * \frac{p(1-p)}{\sqrt{n}}$$

$$IC_{95\%} = p + z * \sqrt{\frac{\sigma^2}{n}} = p + 1,96 * \frac{p(1-p)}{\sqrt{n}}$$

avec :

- p la proportion de remplacements résiliés
- z le coefficient dépendant du niveau de confiance choisi, ici 95%
- σ^2 la variance de l'échantillon
- n la taille de l'échantillon

Pour rappel, plus l'intervalle de confiance est resserré, plus le taux de résiliation est précis. A l'inverse, plus l'intervalle de confiance est grand, moins la précision du taux de résiliation est élevée.

Pour des raisons de confidentialité, les volumes exacts des modalités de chacune des variables ont été masqués. Seuls la proportion de remplacements résiliés dans les 180 jours et les intervalles de confiance sont disponibles. Néanmoins, sur chacune des représentations graphiques, l'histogramme représente les volumes de remplacements dans chaque classe. L'axe des ordonnées indiquant le volume exact a volontairement été caché lui aussi.

2.1 Variables relatives à l'assuré

2.1.1 Age du conducteur principal

La représentation graphique de la figure 2.2.1 indique que près de deux tiers des remplacements sont effectués par des conducteurs âgés de plus de 40 ans. Néanmoins, ce sont ceux qui résilient le moins suite à remplacement

d'après le tableau de cette même figure. Les conducteurs principaux qui ont entre 14 et 20 ans sont ceux qui résilient le plus après remplacement. Du fait d'une faible fréquence des remplacements sur cette tranche d'âge, l'intervalle de confiance est plus grand, ce qui signifie que la précision de cette proportion est plus faible. La représentation graphique permet également de noter que le taux de résiliation suite à remplacement décroît avec l'âge du conducteur. Cela peut s'expliquer par le fait que les plus jeunes sont plus sensibles au prix et cherchent donc à bénéficier du tarif le plus attractif possible.

Classes d'âge du CP	% rpl résiliés	IC 5%	IC 95%
[14;20[	10,9%	9,4%	12,4%
[20;25[	10,2%	9,6%	10,7%
[25;30[	8,7%	8,2%	9,1%
[30;40[	6,6%	6,4%	6,9%
>=40	4,6%	4,4%	4,7%

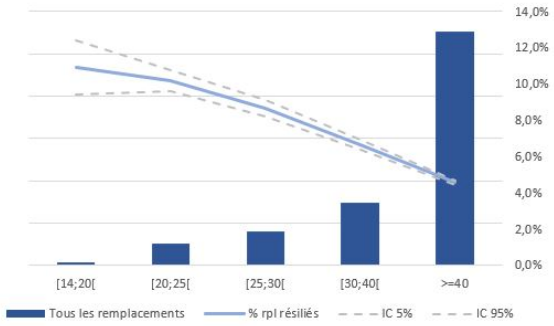


Figure 2.2.1 : Analyse univariée de l'âge du conducteur principal

Source : Base des remplacements

2.1.2 Ancienneté de permis au remplacement

La figure 2.2.2 représente de façon numérique et graphique l'ancienneté de permis du conducteur principal au moment du remplacement. Les individus qui remplacent le plus détiennent leur permis depuis plus de 29 ans. Ceux qui résilient le plus ont eux une ancienneté de permis entre 1 et 2 ans. Le taux de résiliation décroît avec l'ancienneté de permis ce qui est cohérent avec l'analyse effectuée sur l'âge du conducteur principal.

Classes d'ancienneté de permis	% rpl résiliés	IC 5%	IC 95%
[0 ; 1[	7,1%	6,4%	7,7%
[1 ; 2[	9,1%	8,3%	9,8%
[2 ; 3[	8,5%	7,8%	9,2%
[3 ; 5[	8,1%	7,6%	8,6%
[5 ; 10[	6,9%	6,6%	7,3%
[10 ; 15[	6,2%	5,9%	6,6%
[15 ; 20[	5,2%	4,9%	5,5%
[20 ; 29[	4,7%	4,5%	4,9%
>= 29	4,4%	4,2%	4,6%

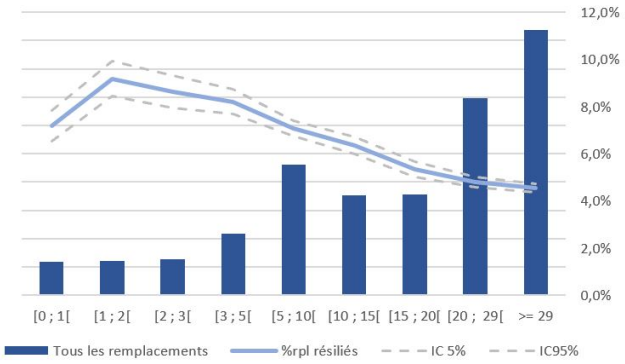


Figure 2.2.2 : Analyse univariée de l'ancienneté de permis au remplacement

Source : Base des remplacements

2.2 Variables relatives aux caractéristiques du contrat

2.2.1 Mouvement de prime

Pour rappel, le mouvement de prime correspond à une modalité qualitative indiquant si la cotisation a augmenté, diminué ou est restée la même au moment du remplacement. Plus de la moitié des remplacements bénéficient d'une baisse de cotisation lors de la modification de leur contrat. C'est également dans ce cas que le taux de résiliation est le plus élevé ce qui est plutôt contre intuitif. En effet, il parait plus intuitif de résilier son contrat en cas de hausse de prime trop importante et non pas en cas de baisse de prime. Ce résultat pourrait néanmoins être expliqué en cas de baisse de prime difficilement compréhensible par le client. Par exemple, si l'assuré change uniquement son adresse et que sa cotisation baisse fortement, il peut considérer qu'il a payé un montant trop important jusqu'à présent et n'aura plus forcément confiance en son assureur. Une autre explication possible vient du fait que les clients dont la prime baisse sont probablement des clients plus sensibles qui ont par exemple volontairement diminué leurs garanties pour baisser le montant de leur prime. En réalité, le taux de résiliation le plus élevé est pour les contrats qui ne changent pas de prime au remplacement. Néanmoins, ce résultat est à prendre avec prudence car l'intervalle de confiance est très large du fait d'un volume très faible de remplacements dans cette modalité.

Mouvement de prime	% rpl résiliés	IC 5%	IC 95%
baisse	6,3%	6,1%	6,4%
hausse	4,9%	4,8%	5,1%
sans changement	6,8%	5,2%	8,5%

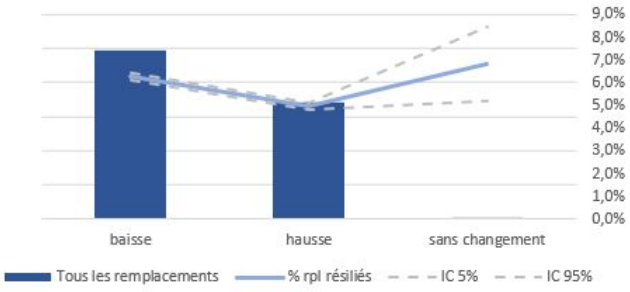


Figure 2.2.3 : Analyse univariée du mouvement de prime au remplacement

Source : Base des remplacements

2.2.2 Taux de crédit commercial agent

Le crédit commercial agent correspond à un avantage tarifaire accordé par les agents généraux pour satisfaire les clients. Une enveloppe avec un budget limité est accordé à chaque agent général en fonction de ses performances. Il peut utiliser ce budget comme il le souhaite : pour attirer de nouveaux clients ou pour fidéliser des clients déjà en portefeuille. Dans le cas du remplacement, l'agent utilise son budget pour fidéliser le client et éviter une résiliation en cas de hausse de prime trop importante. Le taux de crédit commercial agent correspond à la proportion de prime financée par l'agent et donc non payée par l'assuré. La figure 2.2.4 indique que le taux de résiliation suite à remplacement est le plus élevé pour les contrats n'ayant pas bénéficié de crédit commercial agent au moment du remplacement. La proportion de résiliations suite à remplacement semble ensuite se stabiliser peu importe le montant de prime pris en charge par l'agent. Cela peut signifier que lorsque l'agent accorde un geste commercial à l'assuré, ce dernier y est sensible et que le niveau de réduction accordé a été fixé de manière pertinente pour chaque client. Il est également important de noter que seulement 15% des remplacements bénéficient d'un geste commercial d'un agent.

Taux de CCA	% rpl résiliés	IC 5%	IC 95%
0	6,6%	6,4%	6,7%
]0 ; 0.08]	4,8%	4,4%	5,1%
]0.08 ; 0.14]	4,2%	3,9%	4,6%
]0.14 ; 0.24]	3,9%	3,6%	4,2%
]0.24 ; 0.76]	3,8%	3,4%	4,1%

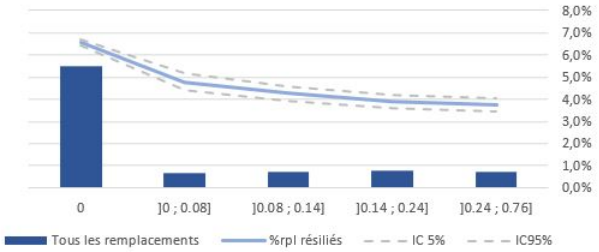


Figure 2.2.4 : Analyse univariée du taux de crédit commercial agent

Source : Base des remplacements

2.2.3 Mois d'effet du remplacement

Comme expliqué lors du contexte de l'étude, la moto est un moyen de transport particulier qui peut être utilisé sur certaines périodes uniquement. Intuitivement, la moto en usage passion est plus utilisée les mois d'été ensoleillés que les mois d'hiver. L'étude du mois d'effet du remplacement peut permettre d'identifier un éventuel effet de saisonnalité. La représentation graphique de la figure 2.2.5 indique que la période qui génère le plus de remplacement s'étend de mars à juillet. La proportion de résiliations suite à des remplacements effectués entre janvier et avril et entre octobre et décembre est la plus élevée sur ces intervalles de temps.

Mois d'effet du remplacement	% rpl résiliés	IC 5%	IC 95%
1	7,2%	6,7%	7,6%
2	6,9%	6,5%	7,4%
3	6,5%	6,2%	6,9%
4	5,4%	5,1%	5,8%
5	5,8%	5,5%	6,2%
6	4,9%	4,6%	5,3%
7	4,5%	4,2%	4,9%
8	4,7%	4,3%	5,1%
9	4,7%	4,3%	5,0%
10	6,3%	5,8%	6,7%
11	6,1%	5,7%	6,6%
12	6,8%	6,2%	7,4%

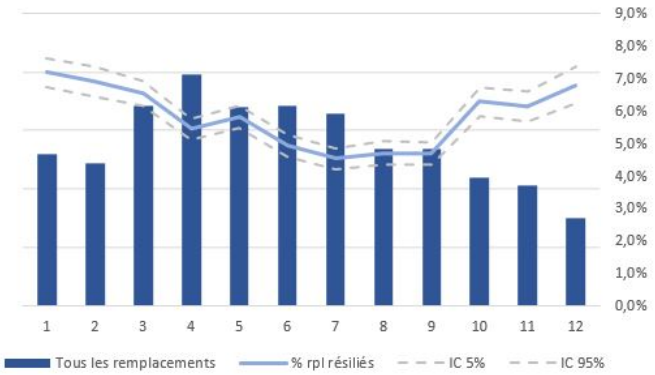


Figure 2.2.5 : Analyse univariée du mois d'effet du remplacement

Source : Base des remplacements

2.3 Variables relatives au comportement assuré

2.3.1 Changement de garanties

La variable changement de garantie permet d'indiquer si, au moment du remplacement, l'assuré a décidé d'augmenter sa couverture ou de la diminuer. Près de deux tiers des remplacements n'ont pas engendré un changement de garantie. Pour le tiers restant, la répartition des remplacements entre diminution et extension de garanties est similaire. Néanmoins, la proportion de résiliations suite à remplacement est beaucoup plus élevée en cas de diminution de garanties. Ce résultat est plutôt contre intuitif car une diminution des garanties engendre dans 96% des cas une baisse de cotisation. Cette analyse rejoint celle effectuée sur le mouvement de prime à la section précédente.

Changement de garanties	% rpl résiliés	IC 5%	IC 95%
sans changement	5,0%	4,9%	5,1%
diminution	9,7%	9,4%	10,0%
extension	4,2%	4,0%	4,5%

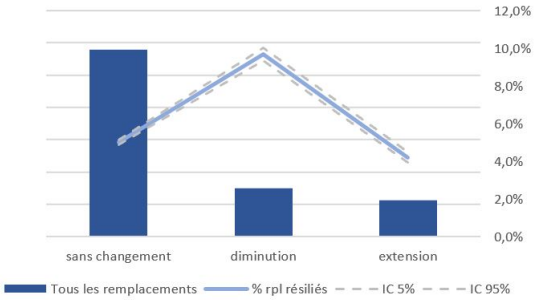


Figure 2.2.6 : Analyse univariée du changement de garanties

Source : Base des remplacements

2.3.2 Graphique de tous les changements

La figure 2.2.7 représente l'ensemble des variables indiquant un changement dans le contrat au moment du remplacement. Pour un même remplacement, un changement de zone et un changement de fractionnement peuvent avoir eu lieu. Dans ce cas, ce dernier apparaît deux fois dans le graphique ci-dessous : une fois dans le nombre de remplacements pour changement de zone et une autre fois dans le nombre de remplacements pour changement de fractionnement. Les remplacements pour changement de véhicule représentent plus d'un tiers de l'ensemble des remplacements de la base, suivis par les remplacements pour changement de garanties et de classe. Les changements qui génèrent le plus de résiliations suite au remplacement sont les changements de garanties, de zone et le retrait d'une clause.

L'augmentation ou la baisse de la cotisation d'assurance suite à un changement de lieu d'habitation peut être difficilement compréhensible pour un client et peut expliquer un rebond des résiliations. De même pour le retrait d'une clause. Si un client assure une automobile et une moto chez AXA, il bénéficie de 25% de réduction sur son contrat moto par le biais d'une clause. Si ce dernier vend son auto, il ne pourra plus bénéficier de la réduction et devra donc payer une cotisation bien plus élevée. La suppression de cet avantage peut engendrer une résiliation de l'assuré.

Tous les top 0/1 sur les changements	% rpl résiliés	IC 5%	IC 95%
chgt véhicule	3,8%	3,6%	3,9%
chgt garanties	7,3%	7,1%	7,6%
chgt classe	4,7%	4,5%	4,9%
chtg genre	4,5%	4,3%	4,8%
chgt marque	4,4%	4,1%	4,6%
chgt groupe	4,4%	4,2%	4,6%
chgt zone	7,5%	7,2%	7,8%
chgt usage	6,8%	6,4%	7,1%
ajout clause	5,1%	4,7%	5,4%
retrait clause	8,3%	7,8%	8,7%
chgt csp	4,3%	3,9%	4,7%
chgt fractionnement	5,8%	5,2%	6,4%
chgt sit mat	5,6%	4,9%	6,3%

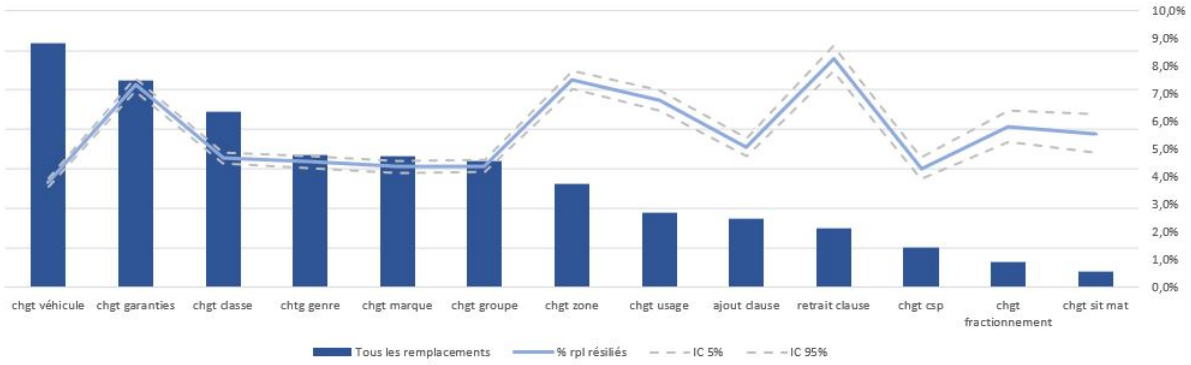


Figure 2.2.7 : Analyse univariée des changements

Source : Base des remplacements

2.3.3 Ancienneté client

L'ancienneté du client n'est pas uniquement relative au contrat moto mais à l'ensemble des contrats détenus par le client. Par exemple, un client assurant son logement depuis 2 ans et sa moto depuis 1 an aura une ancienneté de 2 ans car il détient un contrat chez AXA depuis 2 ans. La figure 2.2.8 met en évidence le fait que les clients qui effectuent le plus de remplacements sont en portefeuille depuis moins de deux ans ou plus de vingt ans. Aussi, la proportion de résiliations suite à remplacement est décroissante avec l'ancienneté client ce qui signifie qu'AXA arrive à fidéliser ses clients.

Ancienneté du client	% rpl résiliés	IC 5%	IC 95%
[0;2]	7,9%	7,6%	8,2%
]2;4]	6,5%	6,2%	6,9%
]4;6]	6,2%	5,8%	6,6%
]6;10]	5,6%	5,3%	5,9%
]10;15]	5,0%	4,7%	5,2%
]15;20]	4,4%	4,1%	4,7%
> 20	3,9%	3,7%	4,2%

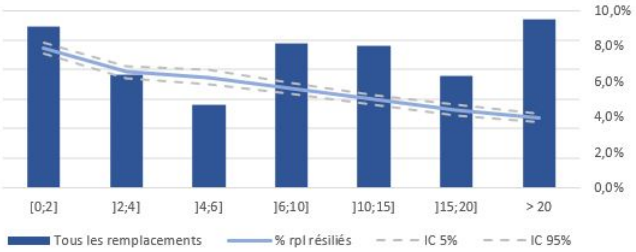


Figure 2.2.8 : Analyse univariée de l'ancienneté du client

Source : Base des remplacements

Chapitre 3

Corrélations et sélection de variables

La corrélation entre deux variables aléatoires permet d'identifier s'il existe un lien entre elles. Afin de mesurer cette dépendance, plusieurs outils sont disponibles et sont à utiliser selon le type de la variable, quantitative ou qualitative.

3.1 Quelques notions théoriques

3.1.1 Corrélation de Pearson et T test

La corrélation de Pearson ([11]*RAKOTOMALALA, 2017*) permet de mesurer l'existence d'un lien entre deux variables aléatoires quantitatives ainsi que l'importance de ce lien. Elle s'exprime sous la forme d'un coefficient compris entre -1 et 1. Lorsque la corrélation vaut -1, les variables sont anti-corrélées. Cela signifie que lorsque l'une des variables augmente, l'autre diminue dans la même proportion. A l'inverse, si le coefficient de corrélation vaut 1, les variables sont parfaitement corrélées et évoluent donc dans le même sens et dans les mêmes proportions. Pour deux variables aléatoires X et Y, le coefficient de corrélation de Pearson est déterminé par

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

avec :

- $Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ la covariance entre X et Y
- σ_X l'écart type de la variable aléatoire X
- σ_Y l'écart type de la variable aléatoire Y

Le calcul de la corrélation de Pearson ([14]*RAKOTOMALALA, 2020*) pour chaque couple de variables quantitatives permet de construire la matrice de corrélations. Cette dernière est carrée, symétrique et les coefficients se situant sur la diagonale sont égaux à 1.

Afin de déterminer la significativité des coefficients de corrélation, un test statistique permettant d'évaluer si le coefficient de corrélation est significativement différent de 0 est initié. Les hypothèses de ce test sont les suivantes :

$$H_0 : \rho_{X,Y} = 0$$

$$H_1 : \rho_{X,Y} \neq 0$$

La p-valeur calculée dans le logiciel R permet d'accepter ou rejeter l'hypothèse nulle en fonction du seuil de confiance fixé. Le seuil de confiance le plus largement utilisé est de 95% ce qui signifie que lorsque la p-valeur est inférieure à 0,05 alors l'hypothèse nulle est rejetée et le coefficient de corrélation est donc significativement différent de 0. A l'inverse, si la p-valeur est supérieure ou égale à 0,05 alors l'hypothèse nulle est acceptée et le coefficient de corrélation n'est donc pas significativement différent de 0.

3.1.2 Test d'indépendance du Chi-deux et V de cramer

Le test d'indépendance du Chi-deux permet d'évaluer s'il existe un lien entre deux variables aléatoires catégorielles. Le nombre d'individus dans chaque échantillon doit être supérieur à 20 pour que le test soit valable. Les hypothèses sont les suivantes :

$$H_0 : \textit{les deux variables ne sont pas corrélées}$$
$$H_1 : \textit{les deux variables sont corrélées}$$

Lorsque la p-valeur est supérieure au seuil fixé de 5%, alors l'hypothèse nulle est acceptée et les variables sont considérées comme non corrélées. A l'inverse, si la p-valeur est inférieure au seuil de 5%, l'hypothèse nulle est rejetée et les variables sont considérées comme corrélées.

Le V de Cramer quant à lui indique l'intensité de la relation entre les deux variables. Il est calculé à partir de la valeur de la statistique du Chi-deux de la façon suivante :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}}$$

Plus la valeur de V est proche de 1, plus l'intensité de la relation entre les variables est élevée. Il faut donc maximiser le V de Cramer pour avoir la relation la plus fiable possible.

3.1.3 Extreme Gradient Boosting

L'Extreme Gradient Boosting ([9] MELLO, 2020), plus communément appelé XGBoost, est un puissant algorithme de machine learning reposant sur la méthode du Boosting. Il permet de développer itérativement des arbres de classification et associe une majoration au poids des observations mal classifiées par le modèle à chaque étape afin de développer le meilleur arbre à la répétition suivante. Il peut être utilisé comme outil de prédiction mais également comme outil d'aide à la décision dans le choix des variables. C'est cette deuxième fonction qui sera utilisée pour déterminer quelle est la variable la plus intéressante à garder dans le cas des corrélations. L'algorithme va ici être appliqué sur l'ensemble de la base, sans séparation en échantillons d'apprentissage et de test. En effet, cet outil étant utilisé ici pour déterminer l'importance des variables dans la base et non prédire le comportement des assurés, toutes les données peuvent être utilisées. Néanmoins, cette méthode ne peut être appliquée que sur des variables quantitatives et non qualitatives. Il est donc nécessaire d'encoder chacune des variables catégorielles afin de leur attribuer une valeur numérique. Pour ce faire, le Label encoding a été utilisé. Cette méthode consiste à attribuer une valeur numérique à chaque catégorie de variables ce qui implique que la colonne contiendra autant de valeurs numériques différentes que de catégories différentes.

Table initiale

Numéro de contrat	Classe d'âge du CP
1454656	>=40
5189553	[30;40[
4617893	[20;25[
7840384	[14;20[
6740485	>=40
5637890	[25;30[
5783906	>=40

Label encoding

Numéro de contrat	Classe d'âge du CP
1454656	1
5189553	2
4617893	3
7840384	4
6740485	1
5637890	5
5783906	1

Figure 2.3.1 : Illustration du Label Encoding

Source : towardsdatascience.com

L'avantage de cette technique est qu'elle n'augmente pas le nombre de colonnes, à la différence du *one hot encoding* ([12] *SHAIKH, 2018*) qui crée autant de colonnes que de modalités dans chacune des variables.

3.2 Les différentes corrélations en pratique

3.2.1 Corrélation des variables quantitatives

Afin d'illustrer la corrélation entre les variables quantitatives, les variables continues montant et évolution de la cotisation totale, taux et montant de crédit commercial agent, montant et évolution de CT, *estimated loss ratio*, ancienneté client, écart de cotisation et délai entre le terme et le remplacement ont été conservées. Néanmoins, comme elles ont été discrétisées pour être utilisées dans le modèle, on retrouvera ces variables continues discrétisées dans l'étude des corrélations des variables qualitatives. Elles ont uniquement été conservées dans le but d'illustrer les différents outils disponibles pour évaluer la corrélation de Pearson.

Afin de pouvoir visualiser plus clairement les variables corrélées, la matrice de corrélations a été représentée à l'aide de la ligne de code suivante :

corrplot(cor(numeriques), order = "hclust", type="upper", tl.col="black")

avec :

- *corrplot* la fonction permettant d'afficher graphiquement les corrélations
- *cor* la fonction permettant de calculer les coefficients de corrélation de Pearson
- *numeriques* l'ensemble des variables continues
- *order = "hclust"* permet de regrouper les corrélations par cluster afin d'obtenir un graphique plus lisible
- *type = "upper"* permet d'afficher uniquement la partie haute de la matrice de corrélations
- *tl.col = "black"* permet d'afficher le nom des variables en noir

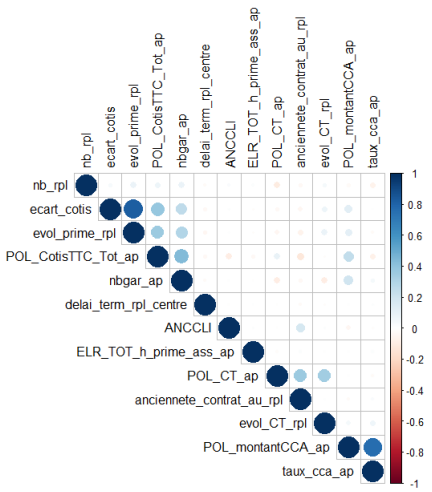


Figure 2.3.2 : Corrélation des variables continues
Source : R (corrplot)

Dans la matrice de corrélation de la figure 2.3.2 ci-dessus, la taille du rond est proportionnelle à la valeur absolue du coefficient de corrélation. Plus la valeur absolue est proche de 1, plus le diamètre du rond est important. La couleur permet d'indiquer le sens de la corrélation. Le bleu indique une corrélation positive entre les deux variables tandis que le rouge indique une corrélation négative. Les variables les plus corrélées, hormis la diagonale, qui correspond à la corrélation des variables avec elles mêmes, sont :

- le montant de crédit commercial agent et le taux de crédit commercial agent
- l'écart de cotisation et l'évolution du montant de prime

La visualisation des corrélations peut également être effectuée avec la fonction *ggpairs* du package R *GGally*. La figure 2.3.3 permet d'obtenir une matrice avec un nuage de points, la représentation de la densité ainsi que la corrélation et sa significativité associée.
Pour rappel, la valeur -100 correspond à la modalité NA (not available) transformée des variables quantitatives.

Les nuages de points permettent d'identifier la distribution des points selon les deux variables. Par exemple, la grande majorité des cotisations inférieures à 1000€ ont un taux de crédit commercial agent entre 0 et 0,5. La densité permet d'avoir une idée de la répartition des valeurs de chacune des variables continues. L'ancienneté du client par exemple est distribuée entre les valeurs 0 et 50 avec un maximum autour de 10 ans. Enfin, l'ensemble des corrélations sont significatives et la plus élevée est celle reliant l'estimated loss ratio et le taux de crédit commercial agent. Elle vaut 0.227.

Le niveau de significativité de la corrélation est hiérarchisé comme suit :

— *** pour les couples de variables les plus significatifs, c'est à dire ceux qui ont une p-valeur inférieure à 0,001

— ** pour les couples de variables qui ont une p-valeur entre 0,001 et 0,01

— * pour les couples de variables qui ont une p-valeur entre 0,01 et 0,05

— . pour les couples de variables qui ont une p-valeur entre 0,05 et 0,1

— pour les couples de variables qui ont une p-valeur supérieure à 0,1, le niveau de significativité est représenté par un blanc.

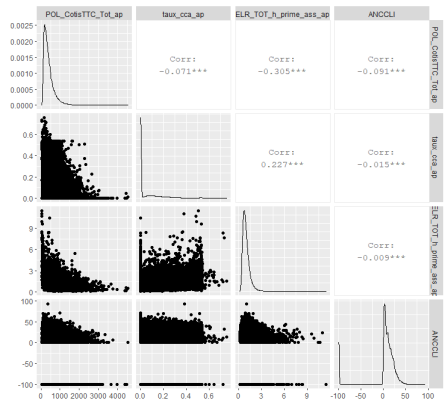


Figure 2.3.3 : Représentation d'un nuage de point, de la densité et des coefficients de corrélations

Source : R (GGpairs)

Les coefficients de corrélation de Pearson ainsi que la p-valeur associée peuvent être extraits de façon numérique. Ce résultat a été mis sous forme d'un tableau à l'aide d'Excel, consultable à l'Annexe 3.3. Dans le cadre de l'étude, l'ensemble des variables continues ont été discrétisées et transformées en variables qualitatives. Le choix entre deux variables corrélées sera donc effectué au niveau des variables discrètes.

3.2.2 Corrélation des variables qualitatives

Le lien entre deux variables qualitatives est mesuré à l'aide du test d'indépendance du Chi-deux et l'intensité de ce lien avec le V de Cramer. Afin de déterminer les variables les plus corrélées, il faut minimiser la p-value du Chi-deux et maximiser le V de Cramer. La figure 2.3.4 recense les 15 variables qualitatives les plus corrélées.

Variable 1	Variable 2	P-value du test du Chi-deux	Valeur du V de Cramer	Place de la variable 1 dans le XGBoost	Place de la variable 2 dans le XGBoost	Variable à supprimer
mouvement_prime	evol_prime_rpl_classe	-	1,00	59	48	mouvement_prime
mouvement_prime	ecart_cotis_classe	-	1,00	59	19	mouvement_prime
POL_OrigineDistrib	NBCVMACT	-	0,99	40	33	POL_OrigineDistrib
POL_OrigineDistrib	ANCCLI_classe	-	0,98	40	56	POL_OrigineDistrib
chgt_genre	chgt_groupe	-	0,84	45	54	chgt_groupe
chgt_marque	chgt_groupe	-	0,83	52	54	chgt_groupe
chgt_gar	type_rpl	-	0,82	5	6	type_rpl
chgt_genre	chgt_marque	-	0,80	45	52	chgt_marque
evol_prime_rpl_classe	ecart_cotis_classe	-	0,75	48	19	evol_prime_rpl_classe
chgt_genre	chgt_csp	-	0,72	45	22	chgt_genre
chgt_usa	chgt_csp	-	0,72	55	22	chgt_usa
chgt_marque	chgt_zone	-	0,72	52	51	chgt_marque
chgt_groupe	chgt_zone	-	0,71	54	51	chgt_groupe
VEH_Genre_ap	VEH_TypeRoues_ap_classe	-	0,71	37	41	VEH_TypeRoues_ap_classe
chgt_genre	chgt_zone	-	0,71	45	51	chgt_zone
chgt_marque	chgt_usa	-	0,71	52	55	chgt_usa
chgt_zone	chgt_usa	-	0,71	51	55	chgt_usa
chgt_groupe	chgt_usa	-	0,71	54	55	chgt_usa
VEH_Cylindree_ap_classe	VEH_TypeRoues_ap_classe	-	0,71	47	41	VEH_Cylindree_ap_classe
chgt_marque	chgt_csp	-	0,71	52	22	chgt_marque
chgt_groupe	chgt_csp	-	0,71	54	22	chgt_groupe
chgt_genre	chgt_usa	-	0,71	45	55	chgt_usa
chgt_zone	chgt_csp	-	0,71	51	22	chgt_zone
annee_effet_an	anciennete_contrat_au_rpl_classe	-	0,68	28	60	anciennete_contrat_au_rpl_classe

Figure 2.3.4 : Variables qualitatives les plus corrélées entre elles

Source : R

Lorsque deux variables sont très corrélées, il est nécessaire d'en supprimer une des deux. Le choix de la variable à conserver peut être effectué à dire d'expert mais il est souvent préférable d'utiliser un autre outil. Dans le cadre de ce mémoire, le choix entre deux variables corrélées a été effectué à l'aide de l'importance de celles-ci dans l'algorithme XGBoost, *Extreme Gradient Boosting*. Après avoir label encodé les variables qualitatives à l'aide de la fonction *LabelEncoder.fit* sur R, les 15 variables qualitatives les plus importantes sont listées dans la figure ci-dessous.

Variable	Importance d'après l'XGBoost
nb_rpl	30,9%
ELR_TOT_h_prime_ass_ap	6,0%
PER_DurDetTarifaire_ap_classe	5,9%
PER_Age_CP_classe	5,7%
chgt_gar	4,1%
type_rpl	3,7%
anciennete_contrat_au_rpl	3,6%
evol_CT_rpl	3,5%
nbgar_ap	3,4%
delai_term_rpl_centre	3,1%
POL_Formule_ap	3,0%
taux_cca_ap	2,6%
POL_Fract_ap	2,4%
evol_prime_rpl	2,3%
mois_effet_rpl	1,9%

Figure 2.3.5 : 15 Variables les plus importantes de l'XGBoost

Source : sorties de l'XGBoost (R et Excel)

Grâce à la figure 2.3.5, les colonnes *place de la variable 1 dans l'XGBoost* et *place de la variable 2 dans l'XGBoost* ont pu être renseignées dans le tableau de la figure 2.3.4. Elles indiquent quelle est la place de la variable en question en terme d'importance. Plus la valeur renseignée dans cette colonne est élevée, plus la variable apparaît loin dans l'XGBoost et donc moins elle a d'importance. La dernière colonne, *variables à supprimer* indique quelle est la variable la plus éloignée dans l'XGBoost et qu'il faut supprimer.

Cette étape a été effectuée une seconde fois afin de vérifier que les variables les plus corrélées aient été supprimées. Dans ce cas, l'importance des variables a été modifiée comme l'illustre la figure 2.3.6. En effet, la suppression de certaines d'entre elles permet de donner plus ou moins de poids aux variables restantes. Par exemple, la variable *ELR_TOT_h_prime_ass_ap_classe* était en deuxième position dans le premier XGBoost avec l'ensemble des variables qualitatives et elle redescend en 6ème position dans le deuxième XGBoost.

Variables	Importance d'après l'XGBoost
nb_rpl	24,5%
PER_DurDetTarifaire_ap_classe	5,3%
PER_Age_CP_classe	5,2%
remplacements	5,1%
chgt_gar	4,8%
ELR_TOT_h_prime_ass_ap_classe	4,6%
delai_term_rpl_centre_classe	4,5%
POL_Formule_ap	4,0%
mois_effet_rpl	3,3%
evol_CT_rpl_classe	2,9%
VEH_DurDetAct_ap_classe	2,9%
ecart_cotis_classe	2,7%
nbgar_ap	2,5%
annee_effet_an	2,5%
POL_Fract_ap	2,5%

Figure 2.3.6 : 15 variables les plus importantes du deuxième XGBoost

Source : sorties de l'XGBoost (R et Excel)

Suite à ces deux étapes successives, 13 variables corrélées ont été supprimées de la base.

3.2.3 Corrélation entre une variable quantitative et une variable qualitative

Il est également possible d'étudier la corrélation entre une variable quantitative et une variable qualitative. Une représentation graphique à l'aide de boxplot, aussi appelées "boîtes à moustaches" peut être tracée comme sur la figure 2.3.7. Cette méthode est assez longue et n'a pas été menée du fait de la discrétisation des variables continues. Le graphique indique que les trois quarts des remplacements qui ont bénéficié d'une baisse de prime ont un montant de cotisation inférieur à 400€. Ceux qui ont subi une hausse de prime ont, pour les trois quarts, une cotisation inférieure à 500€. Enfin, les contrats n'ayant pas changé de prime au remplacement ont une prime inférieure à 600€ dans 75% des cas.

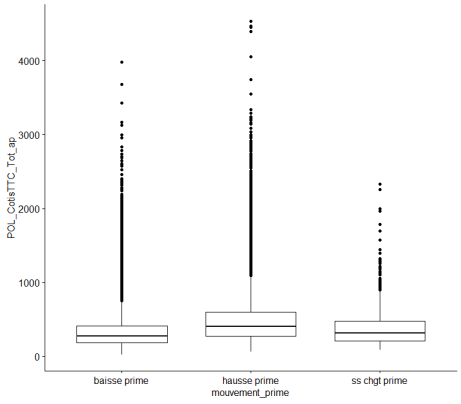


Figure 2.3.7 : Boxplot entre le mouvement de prime et la cotisation totale

Source : R (GGplot

3.2.4 Corrélation entre la variable cible et les autres variables

Il est également possible de déterminer les corrélations entre la variable cible et l'ensemble des autres variables de la base. Jusqu'à présent, les analyses menées ont permis de supprimer les variables trop corrélées entre elles. Le but de cette dernière étape est d'identifier d'éventuelles variables qualitatives complètement décorrélées de l'indicateur de résiliation. Autrement dit, si une variable est très peu corrélée avec la variable cible, alors elle a très peu de chance d'influer sur la prédiction des résiliations.

La figure 2.3.8 ci-contre indique que le changement de fractionnement ne joue pas sur la résiliation dans les 180 jours qui suivent le remplacement. En effet, la p-valeur du test du Chi-deux est bien plus élevée que le seuil de 5% fixé. Cela signifie que l'hypothèse nulle est acceptée, c'est à dire que les deux variables ne sont pas corrélées. Le V de Cramer indique que cette relation entre les variables est fiable à 90%. La variable changement de fractionnement devra donc être écartée pour la suite de l'étude.

Variable 1	Variable 2	P-value du test du Chi-deux	Valeur du V de Cramer
chgt_fract	top_rsl_180j_categ	0,80	0,90
CLA_sinistralite	top_rsl_180j_categ	0,04	0,94
REG_ZoneGarage_RC_ap	top_rsl_180j_categ	0,03	0,98
mois_effet_an	top_rsl_180j_categ	0,00	1,00
chgt_incomp_cli_baisse	top_rsl_180j_categ	0,00	0,88
chgt_incomp_cli_hausse	top_rsl_180j_categ	0,00	0,87
chgt_mois_echeance	top_rsl_180j_categ	0,00	0,94
region_ap	top_rsl_180j_categ	0,00	0,99
VEH_Usage_ap	top_rsl_180j_categ	0,00	0,96
POL_CotisTTC_Tot_ap_classe	top_rsl_180j_categ	0,00	0,92

Figure 2.3.8 : 10 plus faibles corrélations entre la cible et les variables quantitatives

Source : Excel

Partie III

Modélisation et mesures de qualité

Cette troisième partie du mémoire est entièrement dédiée à la modélisation et aux mesures de qualité de celle-ci. Dans un premier temps, une explication des principes mêmes de la modélisation actuarielle supervisée sera énoncée avant d'aborder les aspects plus théoriques liés notamment aux modèles linéaires généralisés. Ensuite, les méthodes de prédictions et de validation des modèles seront abordées avant une mise en pratique sur la base de données constituée. Enfin, les différentes mesures de qualité des modèles et des prédictions permettront de choisir le modèle le plus performant.

L'ensemble des étapes de cette partie ont été effectué sur R et mises en forme sur Excel.

Chapitre 1

Théorie de la modélisation

1.1 Les grands principes de la modélisation

La modélisation permet de traduire une situation réelle en langage mathématique. Il existe une multitude de modèles permettant cette traduction et il est donc important de choisir le plus adéquat. Les modèles descriptifs permettent de représenter un ensemble de données et donc de passer du monde réel au modèle. Les modèles prédictifs permettent quant à eux d'anticiper des évènements futurs en passant du modèle vers le réel. Ces deux types de modèles sont complémentaires et sont utilisés dans le cadre des analyses actuarielles. Une loi de probabilité est déduite de la distribution des observations historiques, ce qui constitue une analyse descriptive des données à l'aide d'une loi théorique. Cette loi est ensuite calibrée et utilisée sur de nouvelles données afin de prédire quels sont les individus qui auront le même comportement. *Prenons un exemple simplifié. Considérons que nous souhaitons déterminer quels sont les élèves de seconde les plus susceptibles de poursuivre en première scientifique. Nous observons les élèves qui sont passés en première S l'année précédente et déterminons que ces derniers avaient tous plus de 13 de moyenne en mathématiques et avaient participé à la fête de la science. Nous avons donc défini une règle permettant de repérer ces individus. Maintenant, nous allons appliquer cette règle sur les élèves de seconde de cette année afin de prédire quels sont ceux qui sont les plus susceptibles de passer en première S. Dans ce cas on entre dans le modèle prédictif qui permet, grâce à la règle obtenue, d'identifier quels seront les futurs élèves de première S.* Cette technique s'appelle l'apprentissage supervisé. Elle permet, à partir d'un échantillon train, d'apprendre comment reconnaître les données que l'on souhaite cibler. Puis cette règle d'apprentissage est appliquée sur un échantillon test et permet de vérifier que le modèle a bien appris.

La variable à prédire peut être continue ou discrète. Les variables continues prennent un nombre infini de valeurs et doivent être modélisées par des lois de probabilité continues ([5] DUTANG, 2020). Les variables aléatoires discrètes quant à elles prennent un nombre fini ou dénombrable de valeurs et doivent être modélisées par des lois discrètes. L'objectif de cette étude est de déterminer quels sont les contrats qui sont les plus susceptibles de résilier suite à remplacement. Cette variable cible prend la valeur 1 si le contrat est résilié dans les 180 jours qui suivent le remplacement et 0 sinon. Le nombre de valeur étant fini, une loi de probabilité discrète doit être utilisée pour modéliser ce comportement. Aussi il est important de noter que les observations sont considérées comme indépendantes les unes des autres dans la mesure où les remplacements n'ont aucun lien entre eux.

1.1.1 Modèle linéaire généralisé et régression logistique

Du modèle linéaire au modèle linéaire généralisé

La variable cible notée *top_ rsl_ 180j* dans la base de données est une variable binaire, c'est à dire qu'elle ne peut prendre que deux valeurs : 0 ou 1. La modélisation d'une variable binaire s'effectue à l'aide de la loi de Bernoulli définie de la façon suivante :

La variable aléatoire X suit une loi de Bernoulli de paramètre p, $X \sim B(p)$, avec $x \in \{0,1\}$ si

$$\mathbb{P}(X = x) = p^x(1-p)^{1-x}$$

$$\textit{Autrement dit, } \mathbb{P}(X = x) = \begin{cases} 1-p & \text{si x} = 0 \\ p & \text{si x} = 1 \\ 0 & \text{sinon.} \end{cases}$$

Le **modèle linéaire** classique permet d'expliquer une variable quantitative Y à l'aide de p variables explicatives X et s'exprime de la façon suivante :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

avec :

- $\beta = (\beta_0, \beta_1, ..., \beta_p)' \in \mathbb{R}^{p+1}$ le vecteur des coefficients linéaires
- $X = (1, X_1, ..., X_p)$ le vecteur des variables explicatives
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ l'erreur standard

Lorsque la variable cible Y est qualitative, ce qui est le cas dans cette étude, la modélisation linéaire est impossible ([15] ROUVIERE, 2017). En effet, la probabilité conditionnelle de Y = 1 sachant X, notée $p(x) = \mathbb{P}(Y = 1|X = x)$ est de la forme

$$p(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

Or $p(x)$ est une probabilité et prend ses valeurs dans l'intervalle [0,1] alors que les valeurs de β ne sont pas contraintes et appartiennent à l'espace des réels tout entier. Le recours aux modèles linéaires généralisés est donc nécessaire pour valider cette égalité à l'aide d'une fonction de lien.

Les **modèles linéaires généralisés** se caractérisent selon trois composantes :

- une loi de probabilité : les variables doivent être indépendantes et appartenir à la famille exponentielle, $(Y_i)_i$ indépendants et $Y_i \sim F(\theta_i, \phi_i, a, b, c)$.
- une fonction déterministe : le prédicteur linéaire est déterminé par les variables explicatives et exprimé par $\eta_i = X_i^T \beta$ avec β le vecteur des coefficients linéaires.
- une fonction de lien g : $\mathbb{R} \mapsto \overline{\mathbb{X}}$ monotone, différentiable, et inversible telle que $E[Y_i|X = x] = g^{-1}(\eta_i(x))$.

La vérification de ces hypothèses est nécessaire en théorie pour la loi de Bernoulli. Elle permet de comprendre le lien entre la distribution initiale et la distribution généralisée.

Première hypothèse

L'objet est de démontrer que les Y_i sont des variables aléatoires indépendantes et que la loi qui les caractérise, la loi de Bernoulli, appartient à la famille exponentielle.

La variable cible correspond à la résiliation suite à remplacement, caractérisée par les valeurs 0 et 1. Comme indiqué précédemment, les remplacements sont indépendants les uns des autres et donc les résiliations suite à remplacement le sont aussi.
Maintenant, il faut démontrer que la loi de Bernoulli appartient à la famille exponentielle, c'est-à-dire qu'elle peut s'écrire sous la forme

$$\ln(\mathbb{P}_{Y_i}(x)) = \frac{\theta x - b(\theta)}{a(\phi)} + c(x, \phi)$$

avec $f_X(x)$ la densité de la variable aléatoire X. Dans le cas de la loi de Bernoulli,

$$\begin{aligned} \ln(\mathbb{P}_{Y_i}(x)) &= \ln(p^x (1-p)^{1-x}) \\ &= \ln(p^x) + \ln((1-p)^{1-x}) \\ &= x \ln(p) + (1-x) \ln(1-p) \\ &= x \ln(p) + \ln(1-p) - x \ln(1-p) \\ &= x \ln(\frac{p}{1-p}) + \ln(1-p) \end{aligned}$$

La loi de Bernoulli est donc une loi de la famille exponentielle de paramètres $\theta = \ln(\frac{p}{1-p})$, $\phi = 1$, $a(\phi) = \phi$, $b(\theta) = -\ln(1-p)$ et $c(x, \phi) = 0$.

Afin de visualiser plus clairement que b est fonction de θ, on peut la réécrire de la façon suivante :

$$\begin{aligned} b(\theta) &= -\ln(1-p) \\ &= \ln(1) - \ln(1-p) \qquad \text{car } \ln(1) = 0 \\ &= \ln(\frac{1}{1-p}) \\ &= \ln(\frac{1-p+p}{1-p}) \\ &= \ln(1+\frac{p}{1-p}) \\ &= \ln(1+\exp(\ln(\frac{p}{1-p}))) \\ &= \ln(1+\exp(\theta)) \end{aligned}$$

La première hypothèse est donc vérifiée.

Deuxième hypothèse

Pour la deuxième hypothèse, la relation $X_i^T\beta$ est bien déterministe. En effet, les variables X_i sont des valeurs observées et donc fixées et les coefficients linéaires β ne sont pas aléatoires non plus car ils sont fixés après estimation. La deuxième hypothèse est donc vérifiée.

Troisième hypothèse

La troisième et dernière composante du modèle linéaire généralisé est la fonction de lien. Elle permet de s'assurer que les valeurs prédites par le modèle respectent bien la structure des observations initiales. L'espérance conditionnelle de Y sachant X = x est, dans le cas de la loi de Bernoulli donnée par :

$$\begin{aligned} \mathbb{E}[Y|X=x] &= \sum_{y=0}^{1} y\mathbb{P}(Y=y|X=x) \\ &= \mathbb{P}(Y=1|X=x) \\ &= \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \\ &= \frac{1}{1+\exp(-x'\beta)} \\ &= g^{-1}(\eta(x)) \\ &= g^{-1}(\beta_0 + x_1\beta_1 + ... + x_p\beta_p) \\ &= g^{-1}(x'\beta) \end{aligned}$$

La fonction de lien g peut donc en être déduite. Pour simplifier les calculs, l'espérance de Y sachant que X = x sera notée p(x).

$$\begin{aligned} \mathbb{E}[Y|X=x] = g^{-1}(x'\beta) \qquad &\Leftrightarrow \qquad p(x) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \qquad (1.1) \\ &\Leftrightarrow \qquad p(x)(1+\exp(x'\beta)) = \exp(x'\beta) \\ &\Leftrightarrow \qquad p(x) = (1-p(x))\exp(x'\beta) \\ &\Leftrightarrow \qquad \frac{p(x)}{1-p(x)} = \exp(x'\beta) \\ &\Leftrightarrow \qquad log(\frac{p(x)}{1-p(x)}) = x'\beta = \eta \qquad (1.2) \end{aligned}$$

La fonction de lien est donc donnée par $g(p) = log(\frac{p(x)}{1-p(x)}) = logit(Y)$. Cette fonction étant monotone, différentiable et inversible, la troisième hypothèse est vérifiée.

La régression logistique et estimateurs

La vérification des hypothèses du GLM a permis de définir le modèle logistique de manière précise. La variable binaire Y correspondant à l'indicateur de résiliation suite à remplacement sera donc modélisée grâce au lien logit qui vérifie

$$logit(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

De façon matricielle, on a

$$x'\beta = logit(y) \Leftrightarrow \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} logit(y_1) \\ \vdots \\ logit(y_n) \end{pmatrix}$$

Afin de déterminer la valeur des coefficients β, plusieurs techniques d'estimations peuvent être utilisées comme la méthode des moments ([4] COTTET, 2019) ou l'estimation par maximum de vraisemblance. C'est cette dernière technique qui est sélectionnée ici et sera implémentée sous le logiciel R.

La fonction de vraisemblance est calculée en effectuant le produit des lois de probabilité de chacune des observations. Pour n observations indépendantes, elle est définie par :

$$\begin{aligned} L_n : \{0,1\}^n \times \mathbb{R}^{p+1} &\to \mathbb{R}^+ \\ (y_1, ..., y_n, \beta) &\mapsto \mathbb{B}(p(x_1)) \bigotimes ... \bigotimes \mathbb{B}(p(x_p))(\{y_1, ..., y_n)\} \\ L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y = y_i | X = x_i) &= \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \end{aligned}$$

La vraisemblance n'est pas facile à manipuler et interpréter. Il est donc préférable d'utiliser la log-vraisemblance pour calculer les estimateurs de β. L'équation devient donc

$$\begin{aligned} \mathcal{L}_n = \log(L_n) &= \log(\prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}) \\ &= \sum_{i=1}^{n} \log(p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}) \\ &= \sum_{i=1}^{n} [y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))] \\ &= \sum_{i=1}^{n} \left[y_i \log(\frac{p(x_i)}{1 - p(x_i)}) + \log(1 - p(x_i)) \right] \\ &= \sum_{i=1}^{n} \left[y_i x_i'\beta + \log(1 - \frac{1}{1 + \exp(-x_i'\beta)}) \right] \\ &= \sum_{i=1}^{n} \left[y_i x_i'\beta + \log(\frac{1 + \exp(-x_i'\beta) - 1}{1 + \exp(-x_i'\beta)}) \right] \\ &= \sum_{i=1}^{n} \left[y_i x_i'\beta + \log(\frac{1}{1 + \exp(x_i'\beta)}) \right] \\ &= \sum_{i=1}^{n} [y_i x_i'\beta + \log(1) - \log(1 + \exp(x_i'\beta))] \\ &= \sum_{i=1}^{n} [y_i x_i'\beta - \log(1 + \exp(x_i'\beta))] \end{aligned} \tag{1.3}$$

Maintenant que la log-vraisemblance a été déterminée, il faut la dériver et l'annuler pour avoir accès aux valeurs de β. Le vecteur qui contient l'ensemble des dérivées premières de la log-vraisemblance en chaque β s'appelle le gradient et est déterminé par

$$\nabla \mathcal{L}_n(\beta) = \left[\frac{\partial \mathcal{L}_n}{\partial \beta_0}(\beta), \cdots, \frac{\partial \mathcal{L}_n}{\partial \beta_p}(\beta)\right]'$$

Pour un β_j donné, j allant de 1 à p, la dérivée est définie par

$$\begin{aligned}
\frac{\partial \mathcal{L}_n}{\partial \beta_j}(\beta) &= \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} [y_i x_i' \beta - \log(1 + \exp(x_i' \beta))] \\
&= \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{x_{ij} \exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right] \\
&= \sum_{i=1}^{n} x_{ij} \left[y_i - \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right] \\
&= \sum_{i=1}^{n} x_{ij} \left[y_i - p(x_i) \right]
\end{aligned}$$

Cette équation n'admet pas de solution explicite. Il est nécessaire d'avoir recours à des algorithmes numériques de descente de gradient pour la résoudre comme par exemple l'algorithme de Newton-Raphson ([15] *ROUVIERE, 2017*). Néanmoins, de nombreux logiciels tels que R et Python permettent le calcul des coefficients linéaires. Grâce à l'estimation des coefficients linéaires, il est possible de calculer la probabilité de résiliation de chacun des remplacements.

1.1.2 La procédure step

La procédure step est une méthode de sélection des variables aléatoires en fonction d'un critère de convergence. Le critère le plus souvent utilisé est l'Akaike Information Criterion, AIC. Cette méthode est appliquée sur un modèle linéaire ou linéaire généralisé et permet d'identifier pas à pas les variables qui font varier l'AIC de ce modèle. Lorsque le retrait d'une variable augmente l'AIC alors elle est conservée dans le modèle. Si néanmoins lorsque cette variable est enlevée l'AIC diminue, alors elle est retirée du modèle. Il existe différentes méthodes de sélections step que sont la méthode forward, la méthode backward et la méthode stepwise.

La méthode forward part du modèle nul, sans aucune variable explicative et ajoute une nouvelle variable à chaque itération. Si cette nouvelle variable augmente l'AIC, alors elle n'est pas retenue. A l'inverse, si elle diminue l'AIC alors elle sera conservée à la prochaine itération. Dans la figure 3.1.1 ci dessous, la variable explicative var2, ajoutée à l'itération 2, permet de diminuer l'AIC. Dans ce cas, les deux variables sont conservées. A l'inverse à l'itération 3, la variable explicative var3 augmente l'AIC. Elle ne sera donc pas conservée à l'itération suivante. Finalement dans cet exemple, le modèle contient 4 variables explicatives.

	Nombre de variables	Variable ajoutée	AIC	Décision
Itération 0	-	-	-	-
Itération 1	1	var1	54 678	-
Itération 2	2	var2	53 456	conservation var1 et var2
Itération 3	3	var3	53 567	suppression var3
Itération 4	3	var4	53 456	conservation var1, var2, var4
Itération 5	4	var5	53 440	conservation var1, var2, var4 et var5

Figure 3.1.1 : Illustration de la méthode step forward

Source : Excel

La méthode backward quant à elle part du modèle complet, avec l'ensemble des variables explicatives et enlève une variable à chaque itération. Si le modèle sans la variable en question a un AIC plus important, il faut conserver la variable dans le modèle. Sinon, cette dernière doit être supprimée. La figure 3.1.2 ci dessous illustre la méthode

backward. Lors de la première itération, la variable explicative var1 est retirée du modèle et implique une augmentation de l'AIC. Cela signifie qu'en gardant les 6 variables de départ, l'AIC est plus faible que celui obtenu sans var1. Il faut donc conserver var1 dans le modèle. L'effet inverse est constaté à l'itération 2. Le modèle final contient dans ce cas 4 variables explicatives.

	Nombre de variables	Variable retirée	AIC	Décision
Itération 0	6	-	53 478	-
Itération 1	5	var1	54 678	conservation var1
Itération 2	5	var2	53 456	suppression var2
Itération 3	4	var3	53 456	conservation var3
Itération 4	4	var4	53 450	conservation var4
Itération 5	4	var5	53 460	suppression var5

Figure 3.1.2 : Illustration de la méthode step backward

Source : Excel

La dernière méthode, stepwise, permet de combiner les deux approches précédentes. A partir du modèle sans aucune variable, une nouvelle variable est ajoutée. Si elle augmente la valeur de l'AIC alors elle est supprimée du modèle. Si néanmoins elle fait baisser la valeur de l'AIC, elle est conservée et la procédure backward intervient. La figure 3.1.3 ci dessous illustre ce mécanisme. La variable var4 de l'itération 4 permet de conserver le même niveau d'AIC. Elle est donc incluse dans le modèle et la sélection backward intervient. Dans ce cas, toutes les combinaisons entre les variables du modèle sont testées. Le deuxième tableau indique que garder uniquement var2 et var4 améliore l'AIC par rapport à la conservation de var1, var2 et var4. Cette combinaison de var2 et var4 n'aurait jamais pu être testée dans le cadre de la méthode forward ou de la méthode backward. C'est en ce sens que la méthode stepwise est la plus complète.

	Nombre de variables	Variable ajoutée	AIC	Décision
Itération 0	-	-	-	-
Itération 1	1	var1	54 678	-
Itération 2	2	var2	53 456	conservation var1 et var2
Itération 3	3	var3	53 567	suppression var3
Itération 4	3	var4	53 456	conservation var1, var2, var4

Itération 4	Combinaison	AIC
backward 1	var1, var2, var4	53 456
backward 2	var1, var2	53 456
backward 3	var1,var4	54 678
backward 4	var2,var4	53 400

Figure 3.1.3 : Illustration de la méthode stepwise

Source : Excel

Ces différentes méthodes sont assez coûteuses en temps lorsque le nombre de variables explicatives du modèle est important. Il est donc nécessaire d'effectuer un premier tri des variables en amont et ne pas appliquer directement cette méthode sur la base.

De même que pour le GLM, la méthode step appliquée à un modèle linéaire généralisé permet d'obtenir l'ensemble des coefficients linéaires β associés aux variables explicatives.

1.1.3 Prédictions

Après avoir calculé les estimateurs des coefficients de régression, le modèle est déterminé. Il faut maintenant utiliser ce modèle pour prédire les futures résiliations, à partir de nouveaux contrats. Le modèle est donc appliqué sur l'échantillon test et donne une probabilité de résiliation entre 0 et 1 pour chaque remplacement. Tout l'enjeu est de réussir à binariser la probabilité donnée afin d'avoir 1 si le contrat est résilié, 0 sinon. Une des méthodes pour déterminer ce seuil optimal réside dans l'arbitrage entre la sensibilité et la spécificité. La sensibilité correspond à la proportion de résiliés correctement prédits et la spécificité correspond à la proportion de non résiliés correctement identifiés. Ces deux concepts sont expliqués plus précisément dans la sous section 1.2.3.

La figure 3.1.4 représente graphiquement la sensibilité et la spécificité en fonction du seuil. Le *cutoff* correspond à la valeur du seuil testée. La courbe noire correspond à la sensibilité en fonction du cutoff. La courbe rouge correspond à la spécificité en fonction du cutoff. Lorsque le seuil est nul, tous les contrats seront considérés comme résiliés. En effet pour rappel, lorsque la probabilité est supérieure au seuil fixé alors le contrat est considéré comme résilié et est identifié par un 1. Inversement, lorsque la probabilité est inférieure au seuil fixé alors le contrat est

considéré comme non résilié et est identifié par un 0. Les probabilités ne peuvent prendre leurs valeurs que dans l'intervalle [0,1]. Dans ce cas, la sensibilité vaut 1, ce qui signifie que tous les contrats réellement résiliés ont été identifiés et la spécificité vaut 0 car aucun contrat non résilié n'a été identifié. A l'inverse, si le seuil vaut 1, aucun contrat ne sera topé à 1 et tous seront considérés comme non résiliés. Dans ce cas, la sensibilité vaut 0 et la spécificité vaut 1. L'enjeu est donc de déterminer le meilleur compromis entre spécificité et sensibilité. Ce dernier correspond au croisement entre les deux courbes.

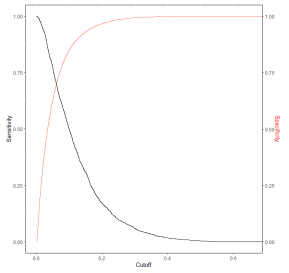


Figure 3.1.4 : détermination du seuil optimal

Source : R

1.2 Stabilité et mesures de qualité

Afin de déterminer quelle modélisation est la meilleure, il est nécessaire de vérifier la stabilité des résultats et d'implémenter des mesures de qualité du modèle mais également des prédictions résultantes.

1.2.1 La validation croisée

La validation croisée est une technique permettant de mesurer la capacité de généralisation d'un modèle et donc la stabilité de celui-ci. Il existe différentes techniques de validation croisée, la plus populaire étant la k-fold ([1] *AZENCOTT, 2020*). A la première itération, une partie des données est utilisée comme test et le reste comme train. A la deuxième itération, une autre partie des données est utilisée comme test et le reste comme train. Et ainsi de suite jusqu'à la k^{me} itération. La figure 3.1.5 illustre cette méthode. Chaque observation de la base aura donc été utilisée une fois dans un échantillon test et une fois dans un échantillon train, ce qui limite l'introduction de biais dans les données.

Figure 3.1.5 : Illustration d'une validation croisée 5-folds

Source : Excel

La validation croisée permet d'identifier la sensibilité du modèle aux données. Autrement dit, si les mesures de qualité des modèles qui seront énoncées ci-après sont très variables d'un fold à l'autre, alors le modèle ne sera pas considéré comme stable. Par exemple, si pour les 5 folds de l'illustration ci-dessus, la précision vaut respectivement 10%, 69%, 90%, 56% et 89% alors la performance dépend beaucoup de la répartition des données en échantillons d'apprentissage et de test et le modèle est donc considéré comme instable. Finalement, la validation croisée est utilisée pour évaluer la capacité de généralisation du modèle.

1.2.2 Mesures de qualité des modèles

Vraisemblance et log-vraisemblance

La vraisemblance d'un modèle, notée L_n permet de mesurer l'adéquation entre une distribution déterminée sur un ensemble d'observations aléatoires et une loi de probabilité qui décrit la réalité. La log-vraisemblance, notée $\mathcal{L}_n$, correspond au logarithme de la vraisemblance et permet de calculer plus facilement les estimateurs des coefficients linéaires comme expliqué précédemment. La valeur de la log-vraisemblance est toujours inférieure à 0. Le meilleur modèle au sens de la vraisemblance est celui qui a une log-vraisemblance la plus proche de 0 possible, autrement dit, celui qui maximise cette valeur. Néanmoins cette valeur est difficile à interpréter car elle dépend de la taille de l'échantillon.

La déviance

La déviance est un outil spécifique qui permet de comparer la vraisemblance obtenue à celle d'un modèle de référence, le modèle complet. Ce dernier reconstitue parfaitement les données à l'échantillon en utilisant autant de paramètres que de données ([15] ROUVIERE, 2017). La déviance d'un modèle est donc définie par

$$D = 2(\mathcal{L}_{sat} - \mathcal{L}_n)$$

avec :

— $\mathcal{L}_{sat}$ la log-vraisemblance du modèle saturé

— $\mathcal{L}_n(\hat{\beta})$ la log-vraisemblance du modèle de régression

Cette mesure de convergence doit être la plus petite possible pour que le modèle soit le meilleur possible. Sur le logiciel R, la déviance est appelée *residual deviance*.

Akaike Information Criterion et Bayesian Information Criterion

L'Akaike Information Criterion et le Bayesian Information Criterion correspondent à deux critères de choix des modèles. Ils permettent de pénaliser la log-vraisemblance pour que cette dernière n'atteigne pas le modèle saturé. En effet, maximiser la vraisemblance revient à choisir le modèle saturé lorsque la complexité du modèle est élevée. Ces critères permettent de favoriser l'utilisation de modèles avec moins de paramètres et donc moins de variables.

L'Akaike Information Criterion (AIC), s'exprime en fonction du nombre de paramètres à estimer p et de la log-vraisemblance du modèle de régression de la façon suivante :

$$AIC = -2\mathcal{L}_n + 2p$$

Il permet d'avoir une estimation de la perte d'information générée par le modèle choisi pour représenter les données.

Le Bayesian Information Criterion (BIC) s'exprime en fonction de la taille de l'échantillon, du nombre de variables et de la log-vraisemblance de la façon suivante :

$$BIC = -2\mathcal{L}_n + p\ln(n)$$

Plus l'échantillon contient d'observations, plus le BIC va être important

Ces deux critères de sélection des modèles doivent être minimisés afin d'obtenir les modèles avec le plus d'informations possible. La valeur de ces indicateurs ne peut être interprétée seule. Elle doit être comparée à celle d'un autre modèle. Le modèle qui minimise ces deux critères sera le meilleur.

1.2.3 Mesures de qualité des prédictions

Matrice de confusion et indicateurs résultants

La matrice de confusion, aussi appelée table de contingence, est un outil permettant de mesurer la qualité de la classification binaire menée. La figure 3.1.6 illustre les différentes catégories permettant de juger de l'efficacité du modèle. Elle compare les prédictions après binarisation avec les valeurs réellement observées.

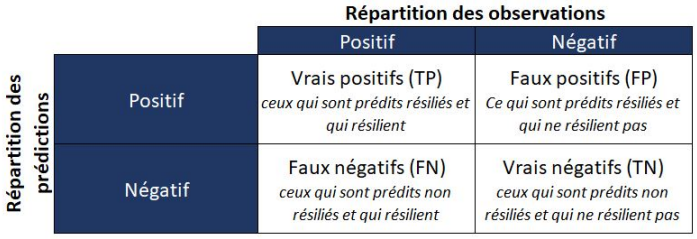


Figure 3.1.6 : Matrice de confusion

Source : Excel

La matrice de confusion permet de déterminer un certain nombre d'indicateurs qui qualifient le modèle. Dans un premier temps, le **taux d'erreur** peut être calculé. Il correspond au nombre de mauvaises prédictions par rapport au total des prédictions, soit

$$taux\_erreur = \frac{FP + FN}{TN + FP + FN + TN} = \frac{FP + FN}{n}$$

Dans le cas idéal où toutes les prédictions sont exactes, l'erreur théorique vaut 0. Dans le cas inverse, si toutes les prédictions sont fausses alors l'erreur théorique vaut 1. La répartition aléatoire des prédictions donne un taux d'erreur de 0,5. Le modèle doit donc surpasser cette valeur pour avoir un intérêt.

Un autre indicateur est la **sensibilité**, aussi appelé rappel. Il correspond à la capacité du modèle à détecter les positifs, c'est à dire les résiliations, et doit donc être maximisé. Autrement dit, il permet de calculer la proportion de positifs correctement identifiés grâce à la formule suivante

$$sensibilite = \frac{TP}{TP + FN}$$

La **spécificité** quant à elle correspond à la capacité du modèle à détecter les négatifs, c'est à dire les non résiliés. Elle doit également être maximisée. Elle permet de calculer la proportion de négatifs correctement identifiés à l'aide de la formule

$$specificite = \frac{TN}{TN + FP}$$

La **précision** permet d'identifier la proportion de positifs correctement prédits parmi les prédictions positives. Autrement dit, c'est la capacité du modèle à repérer les véritables résiliations. Elle doit être maximisée et est calculée par la formule

$$precision = \frac{TP}{TP + FP}$$

Enfin, le dernier indicateur utilisé dans le cadre de ce mémoire pour déterminer la qualité d'une prédiction est la F_k**-mesure**. Elle correspond à une combinaison de la précision et du rappel et doit être maximisée. Elle est déterminée par

$$F_k = \frac{(1 + \beta^2) * rappel * precision}{\beta^2 * precision + rappel}$$

Lorsque k = 1, le même poids est donné à la précision et au rappel. C'est le cas le plus utilisé. Néanmoins, il est possible d'attribuer à k une valeur inférieure à 1 qui signifie que la précision devient plus importante que le rappel. Inversement, si k prend une valeur supérieure à 1 alors le rappel devient plus important que la précision.

La courbe ROC et l'AUC

La courbe ROC, *receiver operating characteristic*, permet de décrire le taux de vrais positifs, qui doit être maximisé, en fonction du taux de faux positifs, de façon graphique. Elle trace l'ensemble de ces couples en fonction de la valeur du seuil de prédictions ([16] *TREMBLAY, 2017*). La courbe ROC du modèle parfait passe par le coin supérieur gauche (0,1). Le meilleur seuil en termes de vrais positifs et faux positifs correspond donc au point le plus proche du coin supérieur gauche. Autrement dit, une partie se superpose avec l'axe des ordonnées. Afin de

définir le meilleur modèle selon ce critère, il est nécessaire de superposer les courbes ROC de chacun des modèles. Celle qui domine les autres en tout point correspond à celle du meilleur modèle. Néanmoins, si les courbes se croisent, il faut faire appel à l'AUC, *area under the curve*, qui correspond à l'aire sous la courbe ROC. Le meilleur modèle correspondra à celui qui a l'AUC le plus élevé. Une ROC-AUC autour de 0.5 indique que la performance du modèle est la même que le modèle aléatoire, c'est à dire autant de chance d'avoir un vrai positif qu'un faux positif. Une classification parfaite donne quant à elle un score de 1 et permet d'identifier uniquement des vrais positifs. L'avantage de l'AUC est qu'elle est indépendante des seuils de classifications. Elle mesure la qualité du modèle sans seuil sélectionné. La ROC-AUC correspond à la mesure de performance de classification la plus communément utilisée.

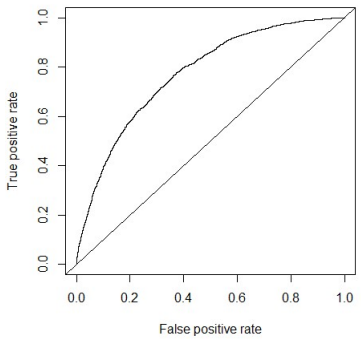


Figure 3.1.7 : Courbe ROC

Source : R

Precision Recall

La courbe Precision Recall ([7] ICHI, 2019) est une autre mesure de performance des modèles en termes de qualité des prédictions. L'objectif est d'identifier l'ensemble des remplacements résiliés, autrement dit avoir un bon rappel, mais surtout que toutes ces prédictions soient justes, donc une bonne précision. Cette courbe permet de s'intéresser aux contrats véritablement résiliés et non pas uniquement aux prédictions des contrats résiliés. Elle permet notamment de se concentrer sur les performances de la classe positive. La courbe Precision Recall du modèle parfait passe par le coin supérieur droit (1,1). Le meilleur seuil en termes de précision et de recall correspond donc au point le plus proche du coin supérieur droit.

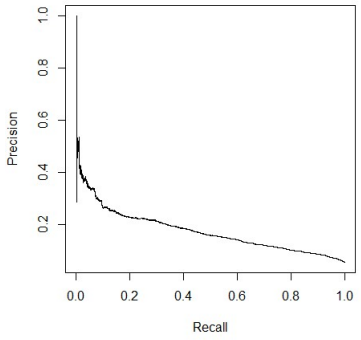


Figure 3.1.8 : Courbe Precision Recall

Source : R

La courbe Lift

La courbe Lift est une mesure de performance permettant de déterminer, pour un pourcentage d'observation fixé, le nombre de contrats résiliés. Autrement dit, elle synthétise les résiliations auxquelles on peut s'attendre par rapport à l'utilisation de l'information à notre disposition. Elle est régulièrement utilisée à des fins de ciblages commerciaux. Dans la figure 3.1.9 ci dessous, 20% des observations, valeur représentée par l'axe vertical rouge, permettent de retrouver près de 55% des contrats résiliés. Pour construire la courbe lift, un vecteur contenant d'une

part les prédictions triées dans l'ordre décroissant et d'autre part le vecteur Y_test a été créé. La taille de la cible correspond au numéro de l'observation classée divisé par le nombre total d'observations. Le rappel correspond au nombre cumulé de vrais positifs divisé par le nombre maximum de vrais positifs.

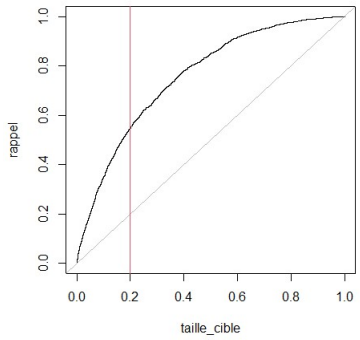


Figure 3.1.9 : Courbe Lift

Source : R

Chapitre 2

Mise en pratique

2.1 Création de la base modèle

2.1.1 Suppression des dernières variables

Avant de pouvoir appliquer un modèle, il est nécessaire de supprimer, en plus des variables corrélées énumérées précédemment, l'ensemble des variables d'identification des contrats telles que le numéro de contrat, le numéro de client et les clés de jointure des tables. Si ces variables sont conservées lors de l'implémentation du modèle, elles seront considérées comme des variables explicatives et un coefficient linéaire leur sera attribué. Cela indiquerait qu'elles jouent un rôle dans la probabilité de résiliation du remplacement. Néanmoins, les résultats associés ne seraient pas exploitables et augmenteraient inutilement le temps d'exécution du modèle. En effet, les numéros de contrat et de client sont définis de façon totalement arbitraire au moment de l'affaire nouvelle et sont uniques pour chaque client. Aucune relation intéressante ne pourrait donc être identifiée.

2.1.2 Changement du type des variables

Suite à la discrétisation des variables effectuée dans la partie 1.3.2 de ce mémoire, les variables qui composent la base sont toutes catégorielles. Elles peuvent être de type caractère ou de type facteur. Le type caractère considère chacune des modalités comme des chaînes de caractères indépendantes. Aucun rapprochement entre elle n'est possible. Le type factor considère la chaîne de caractère comme un niveau et permet donc d'effectuer des comparaisons entre les différentes observations. La figure 3.2.1 ci-dessous permet d'effectuer la comparaison entre les deux types de variables. Elle constitue une sortie du logiciel R obtenue à l'aide de la fonction *summary*.

```
                      ecart_cotis_classe
                     ]-75,-30] :25739
                     ]-150,-75]:25070
                     <= -150   :24338
ecart_cotis_classe   ]-30,0[   :23398
Length:167929        ]80,185]  :17260
Class :character     ]25,80]   :17254
Mode  :character     (Other)   :34870
```

Figure 3.2.1 : Sortie de la fonction *summary*

Source : R

Les variables catégorielles doivent donc être transformées en facteur. Cette modification du type de variable permet de regrouper les différentes modalités et d'identifier clairement les caractéristiques propres à chaque remplacement. Elle se fait à l'aide de la commande suivante sous le logiciel R :

$$base\_modele\$variable <- as.factor(base\_modele\$variable)$$

Afin de vérifier que l'ensemble des variables ont été correctement modifiées, la fonction *summary* sur R est utilisée.

La partie gauche de la figure 3.2.1 illustre la sortie de cette fonction lorsque la variable est de type caractère et la partie droite la même variable lorsqu'elle est de type facteur.

2.2 Modèle GLM

Le premier modèle implémenté pour tenter d'identifier les résiliations consécutives à un remplacement est le modèle linéaire généralisé. La stabilité de celui-ci sera ensuite évaluée à l'aide d'une validation croisée.

2.2.1 Modélisation et prédictions

Séparation train / test et implémentation

L'apprentissage supervisé consiste à entraîner un modèle sur une partie des données et à tester ses capacités de prédiction sur une autre partie des données. Cette technique est utilisée pour que le modèle puisse être testé dans les conditions les plus réalistes possibles. En effet, un modèle qui apprend à reconnaître les remplacements résiliés et qui teste sa capacité d'apprentissage sur ce même jeu de données n'est pas cohérent. En réalité, le modèle est utilisé pour prédire les résiliations suite à remplacement et cette information ne peut donc pas être connue à l'avance. Les données ont donc été divisées en deux sous-échantillons : l'échantillon d'apprentissage qui contient 80% des observations de la base totale et l'échantillon test qui contient les 20% restants. La répartition selon ces proportions est celle la plus utilisée. Aussi, cette dernière doit être effectuée de manière aléatoire afin de limiter le biais dans les résultats. Pour ce faire, un index de répartition est créé sous R, puis les bases d'apprentissage et de test en sont déduites.

$$
\begin{aligned}
&train\_index \leftarrow sample(1 : nrow(base\_modele), 0.8 * nrow(base\_modele)) \\
&test\_index \leftarrow setdiff(1 : nrow(base\_modele), train\_index) \\
&base\_train \leftarrow base\_modele[train\_index,] \\
&base\_test \leftarrow base\_modele[test\_index,]
\end{aligned}
$$

Suite à cette répartition, les deux sous-échantillons ont été générés. La base train contient l'ensemble des observations qui vont permettre au modèle d'apprendre à identifier les remplacements résiliés. L'ensemble des variables explicatives ainsi que la variable cible à prédire la composent. Il en est de même pour la base test qui, elle, contient l'ensemble des observations sur lesquelles le modèle va tester ses connaissances et vérifier qu'il a bien appris.

Le modèle GLM logit est implémenté sur la base d'apprentissage à l'aide de la commande suivante

$$glm\_final \quad \leftarrow \quad glm(top\_rsl\_180j \sim ., data = base\_train, family = binomial(logit))$$

avec :

- *top_rsl_180j* la variable cible que l'on souhaite prédire
- $\sim .$ qui indique que toutes les variables explicatives de la base train sont utilisées pour décrire la variable cible
- *data = base_train* indique que le modèle est entraîné sur la base base_train.
- *family = binomial(logit)* indique que la famille de lois utilisée pour paramétrer le modèle est la famille binomiale avec le lien logit.

La commande glm de R utilisée permet d'avoir accès non seulement aux estimateurs des coefficients linéaires mais également à l'erreur standard de cette estimation, la Z-value ainsi que le niveau de significativité. L'erreur standard correspond à l'erreur susceptible d'être commise sur l'estimation. Plus elle est faible, meilleure sera la confiance apportée à l'estimateur. La Z-value permet de déterminer l'importance de la variable dans la régression. Elle correspond au rapport entre le coefficient et l'erreur standard. Plus elle importante, plus elle indique que le coefficient linéaire est différent de 0 et donc plus la variable explicative associée est importante. Enfin la significativité des coefficients est déterminée à l'aide de la valeur Pr(>|z|). Elle est composée de 5 niveaux :

- *** qui correspond au niveau le plus significatif, c'est à dire $\Pr(>|z|) < 0.001$
- ** qui indique que $0.001 < \Pr(>|z|) < 0.01$
- * qui indique que $0.01 < \Pr(>|z|) < 0.05$

— . qui indique que $0.05 < Pr(>|z|) < 0.1$

— un espace qui indique que $Pr(>|z|) > 0.1$

Les 20 variables les plus significatives d'après le GLM sont recensées dans la figure 3.2.2 ci-dessous. La valeur du coefficient linéaire ainsi que les trois indicateurs qui viennent d'être définis sont également exposés.

Variables et modalités	Coefficient	Valeur absolue du coefficient	Erreur standard	z value	Valeur absolue z value	Pr(>|z|)	Significativité
remplacements1	3,09	3,09	0,11	29,17	29,17	0,00	***
remplacements2	1,78	1,78	0,11	16,33	16,33	0,00	***
PER_DurDetTarifaire_ap_classe[6;8[	-1,02	1,02	0,07	-13,78	13,78	0,00	***
PER_DurDetTarifaire_ap_classe>=10	-1,06	1,06	0,08	-13,61	13,61	0,00	***
(Intercept)	-4,96	4,96	0,39	-12,65	12,65	0,00	***
PER_DurDetTarifaire_ap_classe[8;10[	-1,09	1,09	0,09	-12,56	12,56	0,00	***
PER_DurDetTarifaire_ap_classe[5;6[	-0,92	0,92	0,08	-11,67	11,67	0,00	***
PER_DurDetTarifaire_ap_classe[3;4[	-0,72	0,72	0,07	-11,11	11,11	0,00	***
PER_DurDetTarifaire_ap_classe[4;5[	-0,77	0,77	0,07	-10,74	10,74	0,00	***
VEH_DurDetAct_ap_classe[1;2[	0,51	0,51	0,05	10,40	10,40	0,00	***
VEH_DurDetAct_ap_classe[2;3[	0,56	0,56	0,06	9,93	9,93	0,00	***
chgt_garsans chgt	-0,40	0,40	0,04	-9,69	9,69	0,00	***
VEH_DurDetAct_ap_classe[3;4[	0,62	0,62	0,07	9,33	9,33	0,00	***
delai_term_rpl_centre_classe0	-0,42	0,42	0,05	-9,10	9,10	0,00	***
PER_DurDetTarifaire_ap_classe[2;3[	-0,51	0,51	0,06	-8,91	8,91	0,00	***
POL_Fract_apMensuel	0,26	0,26	0,03	8,47	8,47	0,00	***
VEH_DurDetAct_ap_classe[5;6[	0,68	0,68	0,08	8,21	8,21	0,00	***
VEH_DurDetAct_ap_classe[6;8[	0,59	0,59	0,08	7,78	7,78	0,00	***
VEH_DurDetAct_ap_classe[4;5[	0,57	0,57	0,08	7,48	7,48	0,00	***
evol_CT_rpl_classe> 0.08	0,50	0,50	0,07	7,15	7,15	0,00	***

Figure 3.2.2 : 20 variables les plus significatives d'après le GLM

Source : Excel d'après sortie R

Lorsque le coefficient est positif, en vert dans le tableau de la figure 3.2.2, alors la variable et la modalité influent positivement sur la résiliation suite à remplacement. Autrement dit, le couple variable modalité augmente la probabilité de résiliation des contrats qui le contiennent. A l'inverse, les coefficients négatifs, en rouge dans le tableau, influent négativement sur la probabilité de résiliation des remplacements. Cela signifie que les remplacements qui ont ces caractéristiques diminuent la probabilité de remplacement. Les principales variables qui ont un impact positif sur la résiliation sont le nombre de remplacements, la durée de détention du véhicule, le fractionnement des cotisations et l'évolution du coefficient technique. Elles sont pour la plupart relatives au contrat de l'assuré. Aucune variable relative au conducteur assuré ne ressort dans les 20 premières variables. Celles qui ont un impact négatif sur la résiliation sont la durée de détention tarifaire, le changement de garanties et le délai entre le remplacement et le terme. Cette fois aussi ce sont principalement des variables contrat. L'intercept dans la figure correspond à la valeur de β_0. L'ensemble des informations du tableau sont à prendre en compte et il ne faut pas conclure qu'une résiliation est successive à un remplacement en interprétant seulement une partie des modalités. C'est l'ensemble des caractéristiques des remplacements résiliés qui permettent de définir cette importance du coefficient.

Mesure de la qualité de la modélisation

Afin de déterminer la qualité de la modélisation, la log-vraisemblance a été calculée. Elle vaut -25 618. L'AIC quant à lui vaut 51 690 et le BIC 53 916. La déviance du modèle vaut 59 116. Pour rappel, ces mesures de qualité ne peuvent être interprétées que si elles sont comparées avec celles d'autres modèles. Elles n'indiquent donc pour l'instant rien à propos de la qualité de ce modèle.

Prédictions sur l'échantillon test et mesures de qualité

Le modèle ayant appris à reconnaître les résiliations suite à remplacement, l'application de celui-ci sur l'échantillon test est nécessaire pour prédire les résiliations sur ces nouvelles données. Pour rappel, la base test contient l'ensemble des variables explicatives et la variable à prédire, *top_rsl_180j*. Or pour tester la capacité réelle de prédiction du modèle, cette donnée doit être inconnue. La base test est donc séparée en deux : une base qui contient l'ensemble des variables explicatives dénommée X_test et un vecteur qui contient l'ensemble des valeurs prises par

la variable cible dénommé Y_test. Les prédictions sont déterminées à l'aide de la commande suivante :

$$pred \leftarrow predict(glm\_final, newdata = X\_test, type = "response")$$
$$predictions \leftarrow prediction(pred, Y\_test)$$

La première ligne de commande permet d'obtenir la probabilité de résiliation pour chaque remplacement sous la forme d'un vecteur. La deuxième ligne quant à elle permet de construire un vecteur contenant les probabilités de résiliation déterminées juste avant ainsi que Y_test, qui contient le label indiquant les contrats réellement résiliés. La figure 3.2.3 ci-dessous représente la répartition des prédictions pour chacun des labels, 0 et 1.

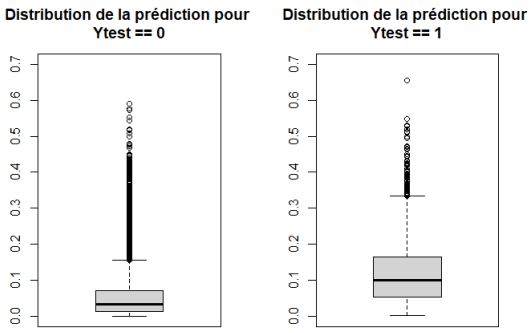


Figure 3.2.3 : Répartition des prédictions du GLM en fonction de la valeur réelle de résiliation
Source : R (GGplot)

Le graphique de gauche sur la figure ci-dessus représente la distribution des prédictions lorsque le contrat n'est en réalité pas résilié. Grâce à la représentation, il est possible de placer graphiquement un seuil de binarisation des prédictions et d'évaluer l'erreur commise. En effet, mettre un seuil à 0.6 permettrait de prédire parfaitement l'ensemble des contrats non résiliés, tous les points étant en dessous de celui-ci. Il faut néanmoins comparer cette valeur avec la distribution des probabilités des résiliés. La boîte à moustache, ou boxplot, est assez resserrée autour de la médiane valant environ 0.05. Les trois quarts des prédictions sont inférieures à 0.1. Cela signifie qu'il est possible de prédire correctement trois quarts des contrats non résiliés en fixant un seuil à 0.1.

Le graphique de droite quant à lui représente la distribution des prédictions lorsque le contrat est en réalité résilié. Le seuil 0.6 identifié comme celui permettant de prédire parfaitement les contrats non résiliés ne permet de prédire correctement qu'un seul contrat résilié. Autrement dit, l'ensemble des autres contrats qui sont réellement résiliés ne seront pas captés en utilisant ce seuil de 0.6. La boîte à moustache est moins resserrée autour de la médiane de 0.1 et les trois quarts des prédictions ont une valeur inférieure à 0.2.

La sélection du seuil de prédiction est difficile dans ce cas. En effet, il n'y a pas de séparation claire entre les probabilités des contrats réellement résiliés et les probabilités des contrats réellement non résiliés. L'enjeu est donc de déterminer le seuil de binarisation des prédictions le plus adéquat à l'aide d'une autre méthode, la représentation de la répartition des probabilités par label ne permettant pas de conclure.

Seuil optimal

Deux nouvelles approches ont donc été menées. La première consiste à trouver le seuil optimal en termes de sensibilité et de spécificité. Pour rappel, la sensibilité correspond à la proportion de positifs correctement identifiés et la spécificité à la proportion de négatifs correctement identifiés. Un arbitrage entre ces deux indicateur permet d'identifier le seuil qui prédit le plus de vrais positifs et de vrais négatifs. Néanmoins, aucune indication n'est donnée sur le nombre de faux positifs et de faux négatifs prédits. Ce seuil optimal ne tient compte que des vrais positifs et des vrais négatifs. Les courbes de sensibilité et de spécificité sont tracées graphiquement en fonction du Cutoff, c'est à dire en fonction de toutes les valeurs possibles que peut prendre le seuil. La figure 3.2.4 ci-dessous représente ces courbes.

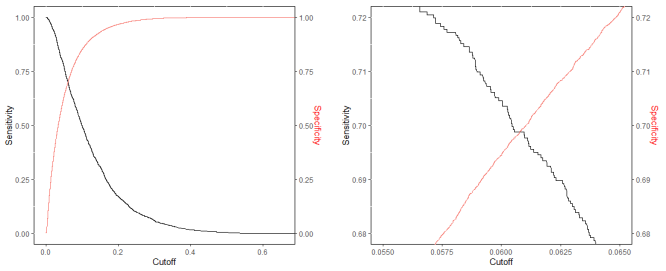


Figure 3.2.4 : Courbes de sensibilité et spécificité pour déterminer le seuil optimal

Source : R (GGplot)

La lecture graphique permet de fixer le seuil optimal de prédiction en termes de sensibilité et de spécificité à 0.06 qui représente le point d'intersection entre les deux courbes. Cela signifie que toutes les prédictions inférieures à 0.06 seront transformées en 0 et toutes celles supérieures à 0.06 seront transformées en 1.

Afin de déterminer l'impact du choix de ce seuil de binarisation sur les données, la matrice de confusion a été déterminée, ainsi que différentes mesures de qualité (Figure 3.2.5). Le vecteur Y_test contient 31 694 contrats non résiliés et 1 892 contrats résiliés. Ces 31 694 contrats non résiliés sont séparés entre les vrais négatifs et les faux positifs alors que les 1 892 contrats résiliés sont séparés entre les faux négatifs et les vrais positifs. Le seuils de 0.06 permet d'identifier correctement 22 010 contrats non résiliés et 1 333 contrats résiliés, soit une sensibilité de 70.5% et une spécificité de 69.4%. Néanmoins, plus de 10 000 contrats sont mal identifiés : 559 sont identifiés comme non résiliés alors qu'ils sont résiliés et 9 684 sont identifiés comme résiliés alors qu'ils ne sont pas résiliés. Ces mauvais classements font chuter la précision du modèle à 12.1% et augmenter le taux d'erreur à 30.5%. La F1-mesure représente le compromis entre la précision et le rappel (ou sensibilité) et doit être maximisée. Dans ce cas, elle est assez faible, autour de 20.7% mais doit être comparée à d'autres modèles. Finalement, tout l'enjeu sera de déterminer s'il est plus avantageux pour AXA de prédire un nombre plus important de faux négatifs ou de faux positifs. Cette étude sera effectuée dans la dernière partie de ce mémoire, une fois la modélisation choisie.

Matrice de confusion

Seuil	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*
0,06	22 010	559	9 684	1 333

Mesures de qualité des prédictions

Specificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
69,4%	70,5%	12,1%	30,5%	20,7%

Figure 3.2.5 : Matrice de confusion et mesures de qualité - GLM

Source : Excel

Pour compléter ces représentations chiffrées de qualité des prédictions, des représentations graphiques sont déterminées notamment les courbes ROC, precision recall et Lift. Ces dernières devront être comparées d'un modèle à l'autre pour pouvoir être exploitées.

La courbe ROC-AUC ci dessous représente la distribution des prédictions en fonction du taux de faux positifs et de vrais positifs pour chaque seuil. Le seuil de 0.06 sélectionné à l'aide de la sensibilité et de la spécificité du modèle est représenté par la croix rouge sur la figure 3.2.6 ci-dessous. Il donne un taux de faux positifs ($\frac{FP}{FP+TN}$) de 30.6% et un taux de vrais positifs (sensibilité) de 70.5%. Pour rappel, le meilleur seuil de prédiction correspond au point de la courbe ROC le plus en haut et à gauche. Le seuil de 0.06 semble correspondre. La courbe ROC est bien meilleure que le modèle aléatoire, représenté par la diagonale. Son AUC vaut 0.77 ce qui constitue une bonne mesure qui devra néanmoins être comparée avec celle des autres modèles.

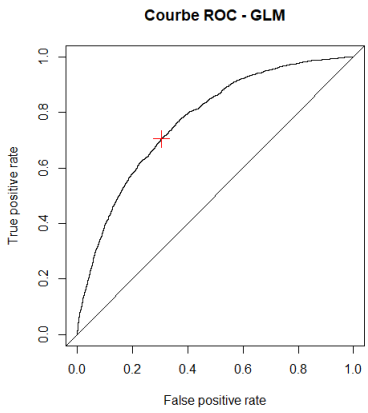


Figure 3.2.6 : Courbe ROC du GLM

Source : R

La courbe precision recall permet d'effectuer un arbitrage entre les contrats prédits comme résiliés et la précision de ces prédictions. Dans le cas de cette étude, moins le recall est important, plus la précision est élevée. C'est à dire que moins les vrais positifs sont détectés, meilleure sera la précision. Et inversement, plus le nombre de vrais positifs est élevé, moins la précision sera importante. Cela s'explique par le fait que plus le nombre de vrais positifs est élevé, plus le modèle a de chances de se tromper et de prédire des faux positifs et des faux négatifs. Dans ce cas, le modèle devient donc moins précis, malgré un nombre important de vrais positifs prédits. La précision et le recall pour le seuil de 0.06 sont représentés par la croix rouge sur la courbe. Pour rappel, plus la courbe est proche du coin en haut à droite, meilleur est le modèle.

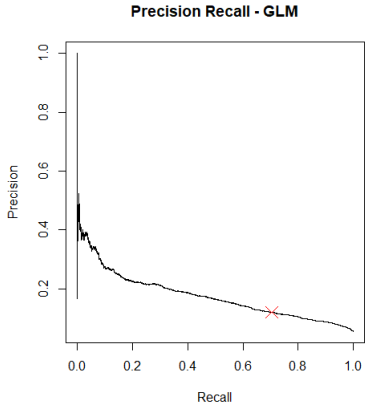


Figure 3.2.7 : Courbe Precision Recall du GLM

Source : R

La courbe Lift quant à elle permet de déterminer, pour une proportion d'observations fixées, la proportion de contrats réellement résiliés. Autrement dit, elle permet de représenter l'amélioration de la sensibilité pour chaque proportion d'observations. La figure 3.2.8 représente la courbe lift du modèle en noir, la courbe lift du modèle aléatoire en gris ainsi qu'un axe vertical rouge qui représente 20% des observations. La lecture graphique permet d'identifier que 20% des observations permettent de retrouver près de 60% des contrats réellement résiliés.

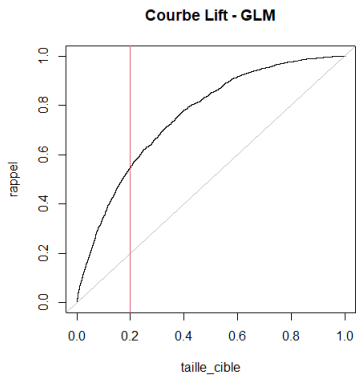


Figure 3.2.8 : Courbe Lift du GLM

Source : R

Seuils arbitraires

La deuxième méthode consiste à fixer des seuils de façon arbitraire et à identifier quels sont ceux les plus performants. Des seuils allant de 0.05 à 0.5 par pas de 5 ont été sélectionnés. A partir du même vecteur de probabilité de résiliations obtenu lors de la prédiction des contrats résiliés à l'aide du GLM, la binarisation a été effectuée pour chacun des seuils. Les remarques effectuées lors de la lecture du graphique 3.2.3 illustrant la répartition des prédictions pour chaque label sont confirmées ici de façon numérique. Plus le seuil de binarisation est élevé, moins le nombre de contrats résiliés prédit sera élevé.

La matrice de confusion pour chacun des seuils ainsi que leurs mesures de qualité respectives ont été déterminées et présentées dans la figure 3.2.9 ci-dessous. Le vecteur Y_test est le même que pour le seuil optimal, seule la distribution des 0 et des 1 des probabilités de résiliation prédites diffère.

Seuil	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	Specificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
0,05	20 037	439	11 657	1 453	63,2%	76,8%	11,1%	36,0%	19,4%
0,10	26 953	951	4 741	941	85,0%	49,7%	16,6%	16,9%	24,8%
0,15	29 499	1 309	2 195	583	93,1%	30,8%	21,0%	10,4%	25,0%
0,20	30 647	1 569	1 047	323	96,7%	17,1%	23,6%	7,8%	19,8%
0,25	31 206	1 708	488	184	98,5%	9,7%	27,4%	6,5%	14,4%
0,30	31 477	1 781	217	111	99,3%	5,9%	33,8%	5,9%	10,0%
0,35	31 592	1 828	102	64	99,7%	3,4%	38,6%	5,7%	6,2%
0,40	31 643	1 860	51	32	99,8%	1,7%	38,6%	5,7%	3,2%
0,45	31 677	1 877	17	15	99,9%	0,8%	46,9%	5,6%	1,6%
0,50	31 685	1 885	9	7	100,0%	0,4%	43,8%	5,6%	0,7%

Figure 3.2.9 : Matrice de confusion et mesures de qualité seuils arbitraires

Source : Excel

Lorsque le seuil est égal à 0.5, la spécificité est maximale et le taux d'erreur très faible ce qui pourrait permettre de sélectionner ce seuil. Néanmoins, la sensibilité et la F1-mesure, qui sont les indicateurs de bonne prédiction des résiliés, sont très faibles. Choisir ce seuil serait donc inutile car il permet de prédire presque parfaitement les contrats qui ne sont pas résiliés et de prédire seulement quelques contrats résiliés. Or le but de l'étude est de prédire les contrats qui sont les plus susceptibles de résilier avec la meilleure précision possible. Les seuils trop élevés sont donc à écarter. Les seuils qui semblent être les plus à même de pouvoir correspondre sont ceux inférieurs à 0.20. La F1-mesure est d'ailleurs maximale pour le seuil de prédiction valant 0.15.

2.2.2 Validation croisée

La validation croisée est implémentée pour déterminer la capacité de généralisation du modèle et donc la stabilité de ce dernier. Le même processus d'implémentation du modèle, de prédiction des valeurs sur l'échantillon test et de mesure de qualité de celles-ci a été suivi.

Choix du nombre folds et implémentation

La première étape pour mener à bien le processus de validation croisée est de déterminer le nombre de *folds* qui séparent les données. Ce dernier ne doit pas être trop faible, pour que la stabilité du modèle puisse être testée plusieurs fois, ni trop élevé, pour que chacune des modalités des variables soit présente dans chaque fold. Les données doivent donc être stratifiées entre chaque fold pour que chaque caractéristique soit représentée dans toutes les subdivisions. La création du nombre de folds dans R se fait à l'aide de la commande suivante :

$$folds\_cv \leftarrow createFolds(factor(base\_modele\$top\_rsl\_180j), k = 10, list = FALSE)$$

Dans le cadre de cette étude et au vu du nombre important d'observations dans la base, 10 folds ont été choisis pour effectuer la validation croisée, permettant ainsi de vérifier la stabilité du modèle tout en respectant la stratification des modalités des variables.

Pour chaque fold, une base train et une base test sont déterminées en fonction des index de cette subdivision. Un GLM est calibré sur l'échantillon train et donne donc des coefficients linéaires propres à chacune des variables qualitatives pour chaque itération. En d'autres termes, la validation croisée 10-folds permet d'implémenter 10 modèles linéaires généralisés sur les mêmes données, avec néanmoins des bases train et test différentes, et de les comparer. La figure 3.2.10 indique les 20 premières variables les plus significatives en moyenne sur les 10 modèles mis en place.

Variables et modalités	Coefficient moyen	Valeur absolue du coefficient moyen	z value moyenne	Valeur absolue z value moyenne	Erreur standard moyenne	Pr(>|z|) moyenne	Significativité
remplacements1	3,15	3,15	30,67	30,67	0,10	0,00	***
remplacements2	1,84	1,84	17,48	17,48	0,11	0,00	***
PER_DurDetTarifaire_ap_classe>=10	-1,06	1,06	-14,34	14,34	0,07	0,00	***
PER_DurDetTarifaire_ap_classe[6;8[	-0,97	0,97	-13,96	13,96	0,07	0,00	***
(Intercept)	-4,97	4,97	-13,29	13,29	0,37	0,00	***
PER_DurDetTarifaire_ap_classe[8;10[	-1,06	1,06	-13,03	13,03	0,08	0,00	***
PER_DurDetTarifaire_ap_classe[5;6[	-0,88	0,88	-11,81	11,81	0,07	0,00	***
PER_DurDetTarifaire_ap_classe[3;4[	-0,69	0,69	-11,36	11,36	0,06	0,00	***
PER_DurDetTarifaire_ap_classe[4;5[	-0,75	0,75	-11,02	11,02	0,07	0,00	***
VEH_DurDetAct_ap_classe[1;2[	0,50	0,50	10,84	10,84	0,05	0,00	***
chgt_garsans chgt	-0,41	0,41	-10,50	10,50	0,04	0,00	***
VEH_DurDetAct_ap_classe[2;3[	0,54	0,54	10,01	10,01	0,05	0,00	***
VEH_DurDetAct_ap_classe[3;4[	0,61	0,61	9,91	9,91	0,06	0,00	***
delai_term_rpl_centre_classe0	-0,42	0,42	-9,63	9,63	0,04	0,00	***
PER_DurDetTarifaire_ap_classe[2;3[	-0,49	0,49	-9,24	9,24	0,05	0,00	***
POL_Fract_apMensuel	0,25	0,25	8,50	8,50	0,03	0,00	***
VEH_DurDetAct_ap_classe[5;6[	0,63	0,63	8,10	8,10	0,08	0,00	***
VEH_DurDetAct_ap_classe[4;5[	0,57	0,57	7,96	7,96	0,07	0,00	***
remplacements3	0,96	0,96	7,94	7,94	0,12	0,00	***
VEH_DurDetAct_ap_classe[6;8[	0,53	0,53	7,34	7,34	0,07	0,00	***

Figure 3.2.10 : 20 variables les plus significatives en moyenne - Validation croisée du GLM

Source : Excel d'après sortie R

Les variables les plus significatives d'après la cross validation sont les mêmes que celles identifiées dans le modèle GLM simple de la figure 3.2.2. Cela indique que la valeur des coefficients du modèle et la significativité de ces derniers sont assez stables. Autrement dit, les 10 modèles donnent en moyenne les mêmes valeurs aux coefficients linéaires et ont sensiblement les mêmes performances en termes de Z-value, d'erreur standard et de significativité.

Mesure de la qualité de la modélisation

Afin de mesurer la qualité de la modélisation, l'AIC, le BIC, la déviance et la log-vraisemblance ont été déterminés dans la figure 3.2.11 ci-dessous. La ligne X1 indique les résultats de la modélisation sur la répartition des données en bases d'apprentissage et de test du premier fold, la ligne X2 indique les résultats de la modélisation sur la

répartition des données en bases d'apprentissage et de test du deuxième fold et ainsi de suite. La dernière ligne *Moyenne* représente la valeur moyenne de ces indicateurs sur les 10 folds. Ce tableau met en lumière la stabilité de ces mesures de performance d'une itération à l'autre. Aucun écart significatif n'est détecté. Ces remarques permettent une nouvelle fois de penser que le modèle est robuste et stable et donc qu'il est possible de le généraliser.

Modèles	AIC	BIC	Deviance	Loglik
X1	57 972	60 225	66 312	- 28 759
X2	58 005	60 259	66 306	- 28 776
X3	58 061	60 315	66 312	- 28 804
X4	58 067	60 320	66 306	- 28 806
X5	58 007	60 260	66 306	- 28 776
X6	57 923	60 176	66 312	- 28 735
X7	58 002	60 255	66 312	- 28 774
X8	58 005	60 258	66 306	- 28 776
X9	57 903	60 156	66 306	- 28 724
X10	57 844	60 097	66 306	- 28 695
Moyenne	**57 979**	**60 232**	**66 308**	**- 28 762**

Figure 3.2.11 : Mesures de qualité de la modélisation

Source : Excel d'après sortie R

Prédictions sur l'échantillon test et mesures de qualité

Le calcul des prédictions et la détermination du seuil optimal ont été effectués de la même façon que pour le GLM classique. Ces étapes ne seront donc pas détaillées pour la validation croisée. Sur la figure 3.2.12 , la matrice de confusion pour chaque fold est représentée ainsi que les mesures de qualité associées. Les seuils optimaux sont compris entre 0,058 et 0,064 avec une moyenne à 0,062. Pour chacun des folds, le seuil optimal est donc relativement proche. La spécificité quant à elle est comprise entre 68,9% et 72,0% avec une moyenne à 70,8%. Là encore, les modèles calibrés à chaque itération permettent de prouver la stabilité des prédictions. La sensibilité est un peu plus volatile avec un minimum à 66,4% et un maximum à 73,1%. Néanmoins, les ordres de grandeur sont assez similaires. La précision minimum vaut 12,4%, son maximum vaut 13,5% et le niveau de précision moyen vaut 12,8%. Il n'y a donc aucun fold pour lequel la précision est bien plus élevée que les autres. Le taux d'erreur est compris entre 28,0% et 30,9% avec une moyenne de 29,2%. Ce dernier est plus de deux fois plus élevé que la précision dans chacun des cas. Il reste néanmoins stable pour chaque itération. Enfin, la F1-mesure est comprise entre 21,1% et 22,5% avec une moyenne de 21,7%. Elle est elle aussi stable d'une itération à l'autre.

Modèles	Seuils optimaux	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	Spécificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
X1	0,061	11 121	290	4 711	671	70,2%	69,8%	12,5%	29,8%	21,2%
X2	0,064	11 135	259	4 696	703	70,3%	73,1%	13,0%	29,5%	22,1%
X3	0,064	11 404	279	4 428	682	72,0%	71,0%	13,3%	28,0%	22,5%
X4	0,063	11 325	283	4 506	679	71,5%	70,6%	13,1%	28,5%	22,1%
X5	0,064	11 175	262	4 656	700	70,6%	72,8%	13,1%	29,3%	22,2%
X6	0,060	11 387	315	4 445	646	71,9%	67,2%	12,7%	28,3%	21,3%
X7	0,064	10 955	273	4 876	689	69,2%	71,6%	12,4%	30,7%	21,1%
X8	0,062	10 914	267	4 917	695	68,9%	72,2%	12,4%	30,9%	21,1%
X9	0,060	11 391	323	4 440	639	72,0%	66,4%	12,6%	28,4%	21,2%
X10	0,058	11 244	283	4 588	678	71,0%	70,6%	12,9%	29,0%	21,8%
Moyenne	**0,062**	**11 205**	**283**	**4 626**	**678**	**70,8%**	**70,5%**	**12,8%**	**29,2%**	**21,7%**

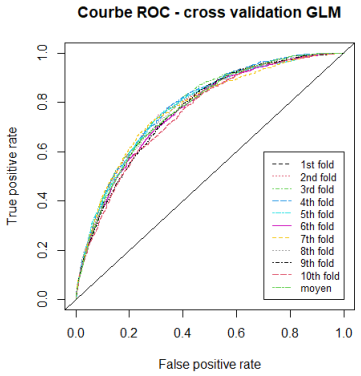
Figure 3.2.12 : Matrice de confusion de la validation croisée du GLM

Source : Excel d'après sortie R

Suite à ces analyses chiffrées permettant de vérifier une nouvelle fois la stabilité du modèle, une représentation graphique de combinaison de ces indicateurs est menée. Les courbes ROC, Precision Recall et Lift sont donc représentées ci-après.

Les courbes ROC obtenues pour chacune des itérations ainsi que celle de la moyenne de ces itérations ont

été représentées graphiquement sur la figure 3.2.13 ci-dessous. La lecture graphique ne permet pas d'identifier un modèle meilleur que l'autre car aucune courbe ne surpasse les autres en tout point. La superposition de ces courbes couvre une surface très faible pour un taux de vrais positifs entre 0 et 0,2 et un taux de faux positifs entre 0 et 0,1 environ. Cela signifie que pour ces taux, et donc pour les seuils qui leur sont associés, les mesures de performance des modèles sont extrêmement proches. Les courbes couvrent une surface plus large lorsque le taux de vrais positifs est compris entre 0,4 et 0,9 et le taux de faux positifs est compris entre 0,2 et 0,7. Les courbes restent néanmoins assez proches et ne permettent pas de rejeter la stabilité du modèle. L'aire sous la courbe ROC est de plus très similaire d'une itération à l'autre avec seulement 2 points d'écart entre le minimum et le maximum. L'AUC moyen sur ces 10 itérations est d'ailleurs le même que celui obtenu pour le modèle linéaire généralisé logit simple.



Modèles	AUC
X1	0,77
X2	0,78
X3	0,78
X4	0,78
X5	0,78
X6	0,77
X7	0,78
X8	0,78
X9	0,77
X10	0,76
Moyenne	**0,77**

Figure 3.2.13 : Courbe ROC pour chaque fold de la validation croisée du GLM et AUC

Source : R

Les courbes Precision Recall pour chaque itération ainsi que la moyenne sur ces dix itérations ont été représentées sur la figure 3.2.14 ci-dessous. De même que pour les courbes ROC, aucune courbe Precision Recall ne surpasse les autres en tout point. Cela indique donc qu'aucun modèle n'est meilleur que les autres et donc qu'ils donnent tous des performances similaires. Les courbes semblent peu stables lorsque le recall est très faible, notamment en dessous de 0,1. En effet, la précision varie de 0,2 à 0,6 pour cette valeur de sensibilité. Néanmoins, le compromis entre la précision et le recall semble se stabiliser pour des valeurs de sensibilité supérieures à 0,4. Une nouvelle fois, la lecture graphique de ces courbes permet de conclure que le modèle linéaire généralisé logit est stable.

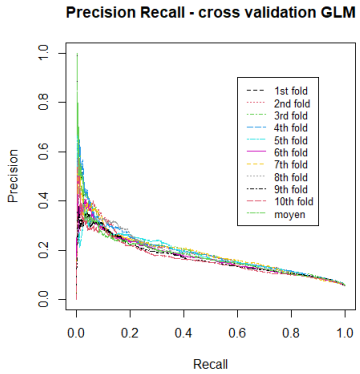


Figure 3.2.14 : Courbe Precision Recall pour chaque fold de la validation croisée du GLM

Source : R

Enfin, les courbes Lift de chaque itération ainsi que celle de la moyenne des dix itérations ont été représentées sur le graphique de la figure 3.2.15 ci-dessous. L'axe vertical en rouge permet de déterminer quelle est la proportion de contrats réellement résiliés identifiée sur chaque fold en utilisant 20% des observations. Le rappel semble prendre des valeurs entre 0,5 et 0,6 pour cette proportion d'observations sélectionnées. Bien que les courbes ne se superposent pas parfaitement, elles sont tout de même très proches et permettent de valider la stabilité du modèle.

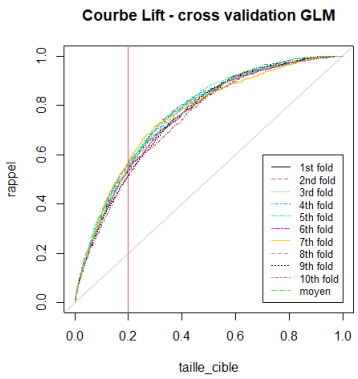


Figure 3.2.15 : Courbe Lift pour chaque fold de la validation croisée du GLM

Source : R

Les différentes analyses menées tant sur la valeur des coefficients linéaires que sur les mesures de qualité et de prédiction des modèles permettent de conclure que le modèle linéaire généralisé implémenté est stable.

2.3 Modèle stepwise

Le deuxième modèle implémenté est la procédure stepwise. Elle permet d'effectuer une sélection de variables en fonction d'un critère de performance, ici l'AIC. Elle est implémentée sur le modèle linéaire généralisé logit simple.

2.3.1 Modélisation et prédictions

Séparation train / test et implémentation

Comme le modèle step prend en entrée le GLM déjà implémenté, les bases d'apprentissage et de test doivent être les mêmes que celles utilisées pour le calibrage du modèle en entrée. La commande R permettant de déterminer pas à pas le meilleur modèle en fonction de l'AIC est la suivante

$$step\_final \leftarrow step(glm\_final, data = base\_train, direction = "backward")$$

avec :

- *glm_final* le modèle linéaire généralisé calibré à la section précédente
- *data = base_train* la base sur laquelle est entraîné le modèle
- *direction = "backward"* indique que la procédure part du modèle avec l'ensemble des variables et élimine, à chaque itération, celle qui fait augmenter l'AIC.

Le choix de la procédure *backward* au profit de la procédure *stepwise* a été motivé par le temps de calcul bien plus faible. La seule procédure *backward* a déjà mis plusieurs heures à tourner sur le logiciel R. C'est également pour cette raison que la direction *forward* a été écartée. Un premier tri des variables ayant déjà été effectué en amont, partir du modèle complet s'est révélé être plus judicieux. En effet, moins de variables risquent d'être retirées du modèle.

La procédure STEP backward a permis d'améliorer l'AIC de 14 points, il passe donc de 51 690 à 51 676. Cette amélioration est due à la suppression des 4 variables suivantes :

- *mois_effet_an* qui correspond au mois d'effet de l'affaire nouvelle
- *VEH_Marque_ap2* qui correspond à la marque du véhicule
- *BEH_SitMat_CP_ap* qui correspond à la situation matrimoniale du conducteur principal
- *chgt_incomp_cli_baisse* qui indique une potentielle incompréhension du client en cas de baisse de prime. *Par exemple, une extension des garanties qui entraîne une baisse de prime est assez contre-intuitif pour un assuré.*

La suppression du mois d'effet de l'affaire nouvelle parait plutôt cohérent car cette variable était peu significative dans le GLM initial. La marque du véhicule ne semble pas indispensable dans le modèle car le genre et le groupe de celui-ci constituent des variables explicatives conservées par le STEP. Enfin, la situation matrimoniale de l'individu et la variable indiquant une incompréhension suite à une baisse de prime sont deux variables non significatives du modèle. Elles peuvent donc être supprimées sans engendrer une perte importante d'information et de performance.

La figure 3.2.16 ci-dessous contient les 20 variables les plus significatives d'après le step. Les coefficients linéaires, l'erreur standard, la Z-value et la significativité du coefficient sont représentés. Les variables qui jouent le plus sur l'augmentation de la probabilité de résiliation suite à remplacement sont le nombre de remplacements, la durée de détention du véhicule, le fractionnement et l'évolution du coefficient technique au moment du remplacement. Ces variables sont principalement liées aux caractéristiques du contrat. Celles qui diminuent la probabilité de résiliation sont la durée de détention tarifaire, le changement de garanties et le délai entre le terme et le remplacement. Là encore, ce sont des variables liées au contrat. Cette sortie permet notamment de noter que lorsque le terme et le remplacement ont lieu le même jour, la probabilité de résiliation diminue. Le nombre de remplacement est la variable qui influe le plus sur la résiliation consécutive à un remplacement, ce qui était déjà le cas dans le GLM logit. Aucun changement majeur suite à la suppression des variables énoncées précédemment n'a eu lieu sur les variables les plus significatives du modèle.

Variables et modalités	Coefficient	Valeur absolue du coefficient	Erreur standard	z value	Valeur absolue z value	Pr(>|z|)	Significativité
remplacements1	3,08	3,08	0,11	29,18	29,18	0,00	***
remplacements2	1,77	1,77	0,11	16,32	16,32	0,00	***
PER_DurDetTarifaire_ap_classe[6;8[	-1,03	1,03	0,07	-13,94	13,94	0,00	***
(Intercept)	-4,91	4,91	0,35	-13,88	13,88	0,00	***
PER_DurDetTarifaire_ap_classe>=10	-1,07	1,07	0,08	-13,82	13,82	0,00	***
PER_DurDetTarifaire_ap_classe[8;10[	-1,11	1,11	0,09	-12,75	12,75	0,00	***
PER_DurDetTarifaire_ap_classe[5;6[	-0,93	0,93	0,08	-11,81	11,81	0,00	***
PER_DurDetTarifaire_ap_classe[3;4[	-0,73	0,73	0,06	-11,24	11,24	0,00	***
PER_DurDetTarifaire_ap_classe[4;5[	-0,78	0,78	0,07	-10,82	10,82	0,00	***
VEH_DurDetAct_ap_classe[1;2[	0,51	0,51	0,05	10,38	10,38	0,00	***
chgt_garsans chgt	-0,41	0,41	0,04	-9,93	9,93	0,00	***
VEH_DurDetAct_ap_classe[2;3[	0,56	0,56	0,06	9,92	9,92	0,00	***
VEH_DurDetAct_ap_classe[3;4[	0,61	0,61	0,07	9,34	9,34	0,00	***
delai_term_rpl_centre_classe0	-0,42	0,42	0,05	-9,18	9,18	0,00	***
PER_DurDetTarifaire_ap_classe[2;3[	-0,51	0,51	0,06	-9,05	9,05	0,00	***
POL_Fract_apMensuel	0,28	0,28	0,03	8,93	8,93	0,00	***
VEH_DurDetAct_ap_classe[5;6[	0,68	0,68	0,08	8,23	8,23	0,00	***
VEH_DurDetAct_ap_classe[6;8[	0,59	0,59	0,08	7,80	7,80	0,00	***
VEH_DurDetAct_ap_classe[4;5[	0,56	0,56	0,08	7,43	7,43	0,00	***
evol_CT_rpl_classe> 0.08	0,51	0,51	0,07	7,28	7,28	0,00	***

Figure 3.2.16 : 20 variables les plus significatives d'après le STEP

Source : Excel d'après sortie R

Mesure de la qualité de la modélisation

L'AIC du modèle step, comme précisé précédemment vaut 51 676. La log-vraisemblance quant à elle vaut -25 665 et le BIC 53 373. La déviance dans le cas du modèle step est la même que pour le GLM logit, elle vaut 59 116.

Prédictions sur l'échantillon test et mesures de qualité

Maintenant que les coefficients du modèle step ont été estimés et que les mesures de qualité de la modélisation ont été déterminées, il est nécessaire d'estimer la probabilité de résiliation des remplacements sur les données de la base test, X_test.

La figure 3.2.17 ci-dessous représente graphiquement la répartition des probabilités de résiliation en fonction du label réel attribué, 0 pour un contrat non résilié et 1 pour un contrat résilié. Les trois quarts des probabilités de résiliation des contrats réellement résiliés ont une valeur inférieure à 0,1. La médiane est située très proche de 0, autour de 0,03. La répartition des probabilités lorsque le contrat est réellement résilié est illustrée sur le graphique de droite. La médiane se situe autour de 0,1 cette fois et les trois quarts de probabilités de résiliation sont inférieures à 0,2. Il est d'usage de définir un seuil de probabilité à 0,5 qui indique que si la probabilité de résiliation est supérieure à 0,5 alors on considère le contrat résilié, sinon non résilié. Néanmoins dans ce cas, la prédiction des contrats résiliés

serait très faible et ne permettrait pas de mener à bien l'étude. Le choix du seuil représente donc un enjeu majeur et plusieurs techniques sont disponibles pour le déterminer.

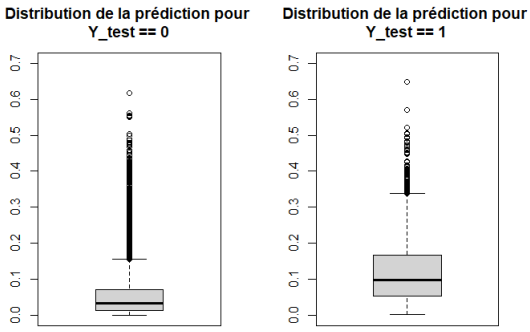


Figure 3.2.17 : Répartition des prédictions du STEP en fonction de la réelle valeur de résiliation

Source : R (GGplot)

Seuil optimal

Le seuil optimal de prédiction est sélectionné à l'aide d'un arbitrage entre la sensibilité et la spécificité du modèle. La représentation graphique de ces deux indicateurs en fonction de la valeur possible des différents seuils, le *cutoff*, est représentée ci-dessous.

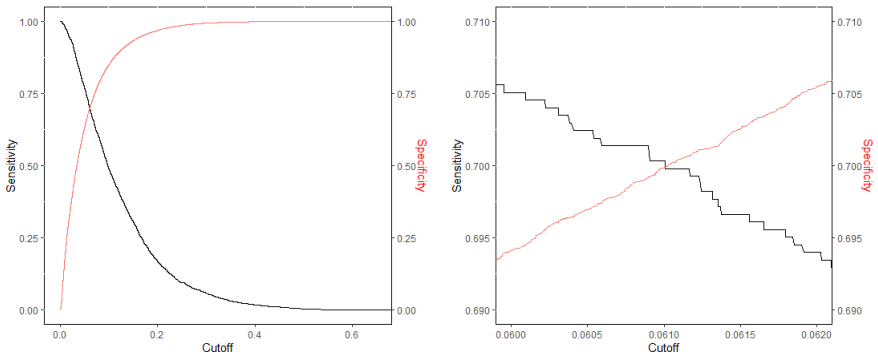


Figure 3.2.18 : Courbes de sensibilité et spécificité pour déterminer le seuil optimal

Source : R (GGplot)

La lecture des graphiques de la figure 3.2.18 ci dessus permet d'identifier un seuil optimal à 0,061, arrondi à 0,06. Les probabilités de résiliation suite à remplacement inférieures à ce seuil seront transformées en 0 et celles supérieures en 1. Autrement dit, lorsque la probabilité de résilier suite à remplacement est inférieure à 0,06 alors le contrat est considéré comme non résilié. Sinon, si la probabilité est supérieure ou égale à 0,06 alors le contrat est considéré comme résilié suite à remplacement.

La matrice de confusion ainsi que les différents indicateurs de performances qui lui sont associés sont déterminés suite à la binarisation des prédictions à l'aide du seuil optimal. Le vecteur Y_test est le même que dans le cas du GLM. Il contient donc toujours 31 694 contrats non résiliés et 1 892 contrats résiliés. La sensibilité et la spécificité du modèle sont les mêmes que dans le cas du modèle linéaire généralisé. Elles indiquent que le modèle step identifie correctement 70,5% des contrats non résiliés et 69,4% des contrats résiliés. Néanmoins, la précision est faible, elle vaut 12,3%, car le nombre de faux positifs est très élevé. Le modèle a tendance à prédire un nombre important de contrats qui vont résilier alors que ce n'est pas le cas. Le taux d'erreur est plus élevé que la précision mais reste relativement acceptable. La F1-mesure quant à elle est assez similaire à celle obtenue dans le cas du GLM.

Matrice de confusion

Seuil	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*
0,06	22 358	579	9 336	1 313

Mesures de qualité des prédictions

Specificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
70,5%	69,4%	12,3%	29,5%	20,9%

Figure 3.2.19 : Matrice de confusion et mesures de qualité - STEP

Source : Excel

La représentation chiffrée des mesures de qualité du modèle est complétée avec une représentation graphique. Les courbes ROC-AUC, Precision Recall et Lift sont tracées.

La courbe ROC-AUC du modèle step ci dessous représente la distribution des prédictions en fonction du taux de faux positifs et de vrais positifs pour chaque seuil. Le seuil de 0,06 est représenté par la croix rouge sur la figure 3.2.20. Il donne un taux de faux positifs de 29,5% et un taux de vrais positifs de 69,4%. Ce point semble bien être celui qui représente le meilleur seuil de prédiction car il est le plus proche du coin en haut à gauche. La courbe ROC du modèle step est également bien meilleure que le modèle aléatoire, représenté par la diagonale. Son AUC vaut 0,77 qui correspond à la même valeur que le modèle GLM logit.

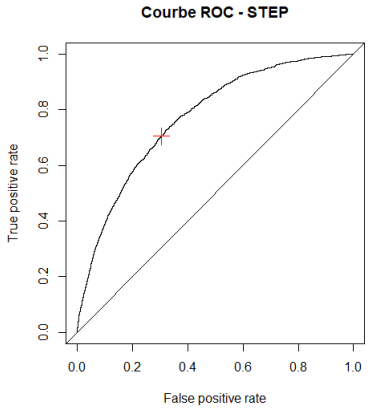


Figure 3.2.20 : Courbe ROC pour le modèle STEP

Source : R

La courbe precision recall du modèle step permet d'effectuer un arbitrage entre les contrats prédits comme résiliés et la précision de ces prédictions. La précision et le recall pour le seuil de 0,06 sont représentés par la croix rouge sur la courbe. Pour rappel, plus la courbe est proche du coin en haut à droite, meilleur est le modèle. Comme pour le modèle GLM, il est clair que le modèle éprouve des difficultés à prédire correctement les contrats résiliés sans faire d'erreur, c'est à dire sans prédire de faux contrats résiliés. C'est pourquoi la courbe est décroissante et un niveau de précision élevé coïncide avec un recall faible.

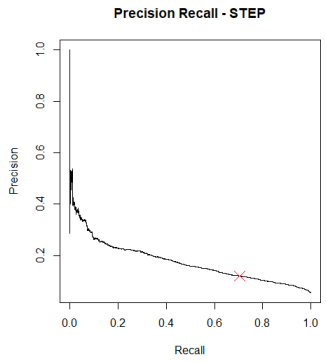


Figure 3.2.21 : Courbe Precision Recall pour le modèle STEP

Source : R

Enfin, la courbe Lift est la dernière mesure utilisée pour analyser la performance du step. La figure 3.2.22 représente la courbe lift du modèle en noir, La lecture graphique permet d'identifier que 20% des observations permettent de retrouver près de 55% des contrats réellement résiliés.

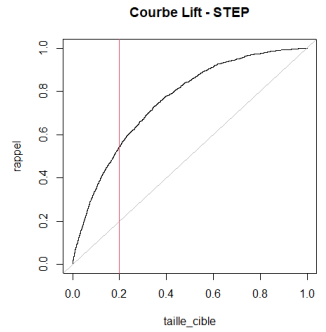


Figure 3.2.22 : Courbe Lift pour le modèle STEP

Source : R

Seuils arbitraires

Comme pour le GLM, différents seuils arbitraires ont été sélectionnés pour binariser les prédictions. Les mêmes seuils, de 0,05 à 0,5 par pas de 0,05 ont été choisis.

La figure 3.2.23 représente la matrice de confusion pour chacun des seuils ainsi que les mesures de performance associées. Le vecteur Y_test qui indique les vraies résiliations est le même que dans le cas du GLM logit.

Seuil	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	Specificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
0,05	19 969	432	11 725	1 460	63,0%	77,2%	11,1%	36,2%	19,4%
0,10	26 891	967	4 803	925	84,8%	48,9%	16,1%	17,2%	24,3%
0,15	29 560	1 315	2 134	577	93,3%	30,5%	21,3%	10,3%	25,1%
0,20	30 675	1 575	1 019	317	96,8%	16,8%	23,7%	7,7%	19,6%
0,25	31 235	1 713	459	179	98,6%	9,5%	28,1%	6,5%	14,2%
0,30	31 477	1 782	217	110	99,3%	5,8%	33,6%	6,0%	9,9%
0,35	31 600	1 836	94	56	99,7%	3,0%	37,3%	5,7%	5,5%
0,40	31 646	1 859	48	33	99,8%	1,7%	40,7%	5,7%	3,3%
0,45	31 675	1 873	19	19	99,9%	1,0%	50,0%	5,6%	2,0%
0,50	31 688	1 887	6	5	100,0%	0,3%	45,5%	5,6%	0,5%

Figure 3.2.23 : Matrice de confusion et mesures de qualité seuils arbitraires

Source : Excel

Il est important de considérer l'ensemble des indicateurs. En effet, maximiser la précision et minimiser l'erreur favoriseraient la sélection du seuil 0,5. Or pour ce seuil, la précision est élevée car le modèle prédit presque parfaitement les contrats non résiliés et presque aucun contrat résilié. Le but de l'étude étant de prédire le nombre de contrats qui résilient dans les 180 jours qui suivent le remplacement, ce seuil n'est pas adapté. Cela est d'ailleurs illustré par la F1-mesure qui est très faible pour ce dernier. Les seuils qui semblent être les plus adaptés au problème sont entre 0,05 et 0,2. La F1-mesure, qui permet un arbitrage entre la précision et le recall, est maximale pour le seuil 0.15, comme dans le cas du GLM.

2.3.2 Validation croisée

Pour des raisons de temps de calcul bien trop important, la validation croisée sur le modèle step n'a pas été effectuée. En effet, la procédure step a mis plusieurs heures à donner des résultats sur R seulement sur le modèle logit. Mener une 10-fold cross-validation aurait pris énormément de temps et n'aurait pas été nécessaire d'un point de vue opérationnel. En effet, les résultats sur ce type d'études doivent être obtenus le plus rapidement possible pour être facilement utilisés par la suite. De plus, le modèle step initial, sans validation croisée, donne des résultats proches des modèles GLM et GLM cross validés. De ce fait, il n'a pas été jugé utile d'effectuer la validation croisée sur le step.

Chapitre 3

Choix du modèle final

Afin de prédire les contrats qui sont les plus susceptibles de résilier dans les 180 jours qui suivent le remplacement, trois modèles ont été implémentés :

— le modèle linéaire généralisé avec le lien logit

— la validation croisée sur ce modèle permettant de justifier la stabilité des coefficients

— le modèle step, issu du modèle linéaire généralisé initial

Un seul des trois modèles doit être retenu et utilisé par la suite pour prédire les résiliations suite à remplacement. Plusieurs critères doivent notamment être pris en compte dans le choix de la modélisation, tant sur les performances que sur le temps de calcul et la mise en place opérationnelle.

3.1 Mesures de qualité de la modélisation

En ce qui concerne les mesures de performance des modèles, le tableau de la figure 3.3.1 ci-dessous indique les valeurs de l'AIC, du BIC, de la déviance et de la log vraisemblance pour chacun des trois modèles. Sans surprise, l'AIC est le meilleur pour le modèle step, bien que celui du GLM ait également une valeur très proche. Le meilleur BIC est également pour le modèle step et assez proche de celui du GLM. Néanmoins, c'est le GLM qui a la meilleure log-vraisemblance, bien que l'écart avec les autres modèles ne soit pas significatif. Enfin, la déviance est la même pour le GLM et le step mais assez éloignée de celle de la validation croisée sur le GLM.

Modèles	AIC	BIC	Déviance	Loglik
GLM	51 690	53 916	59 116 -	**25 618**
Step	**51 676**	**53 373**	**59 116** -	25 665
GLM - crossval 10 flods	57 979	60 232	66 308 -	28 762

Figure 3.3.1 : Comparaison de la qualité de la modélisation

Source : Excel

Les mesures de qualité de la modélisation ne permettent pas d'écarter un modèle en particulier. La validation croisée semble tout de même avoir les moins bonnes mesures de qualité, bien qu'elles ne soient pas significativement éloignées des meilleures d'entre elles.

3.2 Mesures de la performance des modèles

L'analyse des mesures de performance des prédictions effectuées par le modèle sur la base test devraient permettre de déterminer quel est le modèle à retenir.

Les seuils optimaux de chacun des modèles sont identiques comme l'illustre la figure 3.3.2. Le modèle qui a la meilleure performance sur chacun des indicateurs est le modèle de validation croisée sur le modèle linéaire généralisé.

La sensibilité est maximale pour la validation croisée ainsi que pour le modèle linéaire généralisé simple. Pour rappel, la sensibilité correspond à la proportion de contrats résiliés correctement identifiés. Les valeurs de la F1-mesures sont proches d'un modèle à l'autre. Le modèle step prédit mieux les vrais contrats non résiliés que les vrais contrats résiliés alors que le GLM classique prédit mieux les contrats réellement résiliés que les contrats réellement non résiliés.

Modèle	Seuils optimaux	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	Spécificité	Sensibilité / Recall	Precision	Taux d'erreur	F1-mesure
GLM	0,06	22 010	559	9 684	1 333	69,4%	**70,5%**	12,1%	30,5%	20,7%
Step	0,06	22 358	579	9 336	1 313	70,5%	69,4%	12,3%	29,5%	20,9%
GLM - crossval 10 flods	0,06	11 205	283	4 626	678	**70,8%**	**70,5%**	**12,8%**	**29,2%**	**21,7%**

Figure 3.3.2 : Comparaison des matrices de confusion et indicateurs de performance

Source : Excel

Les représentations graphiques de la figure 3.3.3 permettent de visualiser que les modèles présentent relativement les mêmes performances. Aucun modèle ne se détache et propose une performance bien plus élevée que les autres. A l'inverse, aucun modèle ne se détache à cause d'une performance bien moins bonne que les autres.

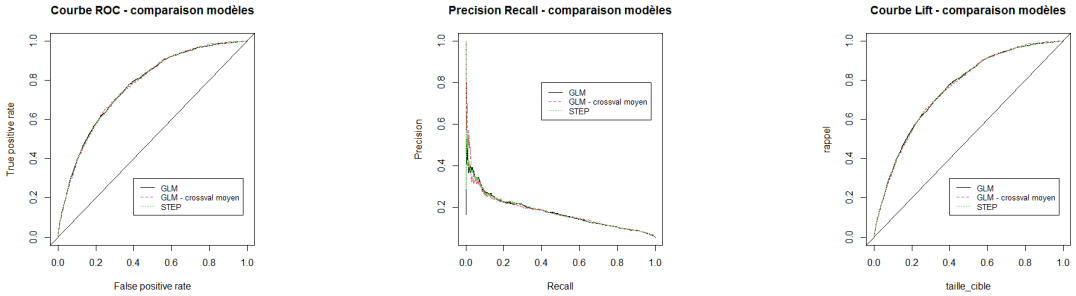


Figure 3.3.3 : Comparaison des courbes ROC, Precision Recall et Lift des 3 modèles

Source : R

Les mesures de qualité des prédictions ne permettent pas de sélectionner de façon certaine un modèle. Néanmoins, la validation croisée du modèle GLM donne les meilleures performances. Le modèle linéaire généralisé pourrait donc être à privilégier.

3.3 Critères opérationnels

Outre les mesures de performance purement mathématiques, l'objectif à terme pour AXA est d'utiliser cette étude sur de nouvelles données. Il faut donc que la modélisation soit la plus performante mais aussi la moins compliquée possible. En effet, les mesures de performance n'étant pas significativement différentes, le choix du modèle à utiliser s'orienterait vers le plus facile à implémenter et à interpréter.

Aussi, dans la même idée, le coût du modèle sélectionné doit être le plus faible possible. Autant le coût temporel que le coût de stockage. Le modèle sélectionné doit permettre d'obtenir des résultats assez rapidement. Aussi, le stockage des résultats sur le serveur ne doit pas être trop important car cette étude peut être amenée à être effectuée tous les mois.

Au vu des différents arguments présentés ci-dessus, le modèle sélectionné pour prédire les résiliations qui interviennent dans les 180 jours qui suivent le remplacement est le modèle linéaire généralisé logit. Le GLM est un outil de modélisation robuste qui a déjà fait ses preuves tant chez AXA que, plus largement, dans le monde de l'actuariat. Aussi, la qualité de la modélisation et les performances des prédictions sont très proches sur chacun des modèles et ne permettent pas de choisir en utilisant ce seul critère. Le GLM est le modèle le plus simple à implémenter et le plus rapide, les résultats sont obtenus en quelques minutes.

Partie IV

Impacts opérationnels

Cette dernière partie démontre l'intérêt de l'étude d'un point de vue purement opérationnel. Elle met en lumière le nombre de clients retenus ainsi que la part de chiffre d'affaire conservée grâce à l'application de mesures tarifaires ciblées.

Les sorties ont été mises en forme sur Excel.

Maintenant qu'il est possible d'identifier les contrats les plus susceptibles de résilier, il faut mesurer l'impact réel de cette prédiction sur la rétention client en termes de volume et de chiffre d'affaire. Les montants prime moyenne en portefeuille et prime au moment de la résiliation présentés par la suite ont été modifiés pour des soucis de confidentialité. Néanmoins, les proportions ont été conservées ce qui ne biaise pas les résultats.

Sans aucune mesure tarifaire

Dans cette section, l'approche purement statistique est menée et la modélisation est simplement appliquée aux données contenues dans le vecteur Y_test. La matrice de confusion de la figure suivante est d'ailleurs bien la même que celle de la figure 3.2.9, avec le seuil optimal en plus. Afin de déterminer l'impact de cette matrice de confusion sur le chiffre d'affaire, les primes moyennes portefeuille et résiliation ont été utilisées. La prime moyenne du portefeuille s'élève à 236€ et la prime moyenne de résiliation s'élève à 286€. Les montants de chiffre d'affaire ont été calculés de la façon suivante :

- **Conservation TN** : les *true negative* correspondent aux contrats qui ne résilient pas et qui sont prédits non résiliés. Dans ce cas, l'intégralité de la prime en portefeuille est conservée : nombre_TN * 236€
- **Perte FN** : les *false negative* correspondent aux contrats qui résilient alors qu'ils étaient prédits non résiliés. Dans ce cas, l'intégralité de la prime de résiliation est perdue : -(nombre_FN * 286€).
- **Conservation FP** : les *false positive* correspondent aux contrats qui ne résilient pas alors qu'ils étaient prédits comme résiliés. Dans ce cas, l'intégralité de la prime portefeuille est conservée : nombre_FP * 236€.
- **Perte TP** : les **true positive** correspondent aux contrats qui résilient et qui ont été prédits comme résiliés. Dans ce cas, l'intégralité de la prime de résiliation est perdue : -(nombre_TP * 286€).

					SANS AUCUNE MESURE APPLIQUEE				
Seuils	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	Conservation TN	Perte FN	Conservation FP	Perte TP	CA total
0,05	20 037	439	11 657	1 453	4 728 732 €	-125 554 €	2 751 052 €	-415 558 €	6 938 672 €
0,06	22 010	559	9 684	1 333	5 194 360 €	-159 874 €	2 285 424 €	-381 238 €	6 938 672 €
0,10	26 953	951	4 741	941	6 360 908 €	-271 986 €	1 118 876 €	-269 126 €	6 938 672 €
0,15	29 499	1 309	2 195	583	6 961 764 €	-374 374 €	518 020 €	-166 738 €	6 938 672 €
0,20	30 647	1 569	1 047	323	7 232 692 €	-448 734 €	247 092 €	-92 378 €	6 938 672 €
0,25	31 206	1 708	488	184	7 364 616 €	-488 488 €	115 168 €	-52 624 €	6 938 672 €
0,30	31 477	1 781	217	111	7 428 572 €	-509 366 €	51 212 €	-31 746 €	6 938 672 €
0,35	31 592	1 828	102	64	7 455 712 €	-522 808 €	24 072 €	-18 304 €	6 938 672 €
0,40	31 643	1 860	51	32	7 467 748 €	-531 960 €	12 036 €	-9 152 €	6 938 672 €
0,45	31 677	1 877	17	15	7 475 772 €	-536 822 €	4 012 €	-4 290 €	6 938 672 €
0,50	31 685	1 885	9	7	7 477 660 €	-539 110 €	2 124 €	-2 002 €	6 938 672 €

Figure 4.3.4 : Chiffre d'affaire sans mesure tarifaire

Source : Excel

La colonne CA total du tableau de la figure 4.3.4 ci-dessus représente la somme par ligne des différents chiffres d'affaires calculés de façon séparée. Le montant est le même pour chaque ligne ce qui est normal. En effet, la somme des TN et FP est la même pour chaque ligne, de même que pour les FN et TP. Comme la prime en portefeuille a été appliquée sur les TN et FP et la prime de résiliation sur les FN et TP, la proportion est conservée. Le chiffre d'affaire total après résiliation et sans aucune mesure tarifaire est donc d'environ 6,9M€. A noter que le chiffre d'affaire avant résiliations vaut 7 926 296€, déterminé de la façon suivante : (20 037 + 439 + 11 657 + 1453) * 236€. Sans appliquer aucune mesure tarifaire, AXA subit une perte de 12,5% de chiffre d'affaire.

Avec mesures tarifaires

Pour rappel, le but de l'étude est d'augmenter la rétention des clients suite à remplacement. Jusqu'à présent, seules les prédictions des contrats les plus susceptibles de résilier ont été déterminées. Il est dorénavant nécessaire d'évaluer les mesures tarifaires à mettre en place ainsi que leurs impacts.

La mesure théorique retenue à ce stade est un avantage de 5% sur la prime des contrats prédits comme résiliés. Le montant de cette réduction a été déterminé de façon à fixer montant de réduction significatif et valorisable auprès des clients concernés afin d'éviter leur départ.

Cette mesure tarifaire ne permettra pas pour autant de conserver 100% des contrats en portefeuille. Certains contrats sont résiliés pour des motifs pour lesquels une baisse de tarif ne fera pas changer d'avis l'assuré. La figure 4.3.5 ci-dessous rappelle la répartition des différents motifs de résiliation dans la base totale, tableau de gauche, et dans la base test, tableau de droite. Il est clair qu'une réduction de la prime pour des contrats résiliés suite à décès (0,6%) ou liquidations judiciaires (0,1%) ne permettra pas de conserver ces contrats. Le motif vente de véhicule est utilisé pour les ventes réelles de véhicule mais également pour les départs vers la concurrence lors d'un changement de véhicule. Une étude annexe à ce mémoire a permis d'identifier que 51% des assurés qui demandent un devis de remplacement partent à la concurrence après avoir vu le tarif. Il serait donc possible d'agir sur l'ensemble des motifs de résiliation autres que décès et liquidation judiciaire ainsi que sur 51% des résiliations suite à vente. Au final, le taux maximum de conservation vaut 64,3% (28,1% + 51% * 70,9%). Cela signifie que dans le meilleur des cas, si tous les contrats sont sensibles à cet avantage tarifaire et qu'ils décident de rester, alors 64,3% seront retenus. Les 35,7% restants étant ceux impossibles à récupérer car résiliés pour des motifs spécifiques (vente réelle de véhicule, décès et liquidation judiciaire).

Le taux de conservation sélectionné, à appliquer sur les 64.3% déterminé précédemment est de 75%. Cela signifie que l'avantage tarifaire permettrait de conserver 75% des contrats qui devaient résilier pour un motif récupérable.

Motifs de résiliation	Proportions base totale
Vente	72,0%
Hamon	15,7%
Loi Chatel	2,8%
Remplace par	2,7%
Echeance	2,2%
Autres cas	1,4%
Sinistre	1,1%
Suite suspension	1,0%
Changement situation	0,4%
Deces	0,4%
Perte Totale	0,3%
Refus majoration	0,0%
Liquidation judiciaire	0,0%

Motifs de résiliation	Proportions base test
Vente	70,9%
Hamon	15,9%
Loi Chatel	2,9%
Remplace par	2,8%
Echeance	2,5%
Sinistre	1,6%
Autres cas	1,1%
Suite suspension	1,0%
Deces	0,6%
Changement situation	0,4%
Perte Totale	0,3%
Liquidation judiciaire	0,1%
Refus majoration	0,0%

Figure 4.3.5 : Détermination du taux de conservation

Source : Excel

Suite à la détermination du taux de conservation, l'ensemble des paramètres de l'étude sont fixés et représentés dans la figure 4.3.6. Le taux de résiliation est calculé de la façon suivante

$$taux\_resil = \frac{Positifs}{Total} = \frac{FN + TP}{TN + FN + FP + TP}$$

Les paramètres qu'il est possible de moduler sont *avantage tarifaire accordé* et *taux de conservation après avantage*. Les autres paramètres sont fixes et dépendent de ces deux valeurs.

Prime moyenne à la resiliation	286
Prime moyenne en portefeuille	236
Avantage tarifaire accordé	-5%
Taux de conservation après avantage	75%
% sur lequel on peut réellement conserver	64%
Taux de conservation réel	48%
CA avant les résiliations	7 926 296 €
CA après si aucune mesure appliquée	6 938 672 €
Taux de résiliation	5,6%

Figure 4.3.6 : Paramètres de calcul

Source : Excel

Prédire un faux positif ou un faux négatif ne pèse pas le même poids en termes de chiffre d'affaire. En effet, un faux négatif est un contrat identifié comme non résilié alors qu'il résilie. Dans ce cas, aucune mesure tarifaire n'est appliquée car le contrat était sensé rester et la totalité de la prime est perdue. Les faux positifs quant à eux représentent des contrats identifiés comme résiliés alors qu'ils restent en réalité. Ils ont donc bénéficié d'un avantage de 5% sur leur prime alors qu'elle n'était pas nécessaire et AXA perd donc 5% de chacune des primes des faux positifs.

Le chiffre d'affaire généré par les différentes prédictions a été calculé de la façon suivante :

— **CA généré par les TN** : dans ce cas, les résultats sont les mêmes que dans l'approche sans aucune mesure tarifaire. En effet, les vrais négatifs qui sont les contrats non résiliés prédits non résiliés n'ont pas bénéficié d'avantage tarifaire et l'intégralité de la prime portefeuille a été conservé.

— **CA généré par les FN** : de même que pour les TN, les résultats obtenus pour les FN sont les mêmes que dans l'approche sans aucune mesure tarifaire. En effet, les contrats sont prédits comme non résiliés mais résilient. Aucune mesure tarifaire n'est donc appliquée et l'intégralité de la prime résiliation est perdue.

— **CA généré par les FP** : les faux positifs correspondent aux contrats prédits résiliés alors qu'ils ne le sont pas. Dans ce cas, une baisse de 5% de la prime leur a été accordée alors qu'elle n'était pas nécessaire. Le taux de conservation est de 100% dans ce cas car les assurés seraient restés même sans cet avantage. Finalement, le chiffre d'affaire généré par les faux positifs vaut CA_FP = nombre_FP * 0,95 * 236€

— **CA généré par les TP** : les vrais positifs correspondent aux contrats qui sont prédits résiliés et qui résilient. Dans ce cas, une remise de 5% est appliquée à chacun des contrats. Néanmoins, l'application de cette réduction ne permet pas de retenir 100% des individus comme expliqué précédemment. Seul 75% * 64% = 48% des individus sont retenus. Le reste d'entre eux résilie. AXA conserve donc 95% de la prime de ceux qui restent après avantage et perd 100% de la prime de ceux qui décident tout de même de résilier. Finalement, le chiffre d'affaire généré par les vrais positifs vaut CA_VP = - 286€* nombre_TP * (1 - taux_conservation_réel) + 236€* nombre_TP * taux_conservation_réel * 0,95.

La figure 4.3.7 ci-dessous indique pour chaque seuil le montant de chiffre d'affaire généré selon la mesure tarifaire appliquée.

					DECOMPOSITION DU CHIFFRE D'AFFAIRE AVEC LES MESURES				
Seuils	TN *True negative*	FN *False negative*	FP *False positive*	TP *True positive*	CA généré par les TN	CA généré par les FN	CA généré par les FP	CA généré par les TP	CA total si mesures
0,05	20 037	439	11 657	1 453	4 728 732 €	-125 554 €	2 613 499 €	-58 056 €	7 158 621 €
0,06	22 010	559	9 684	1 333	5 194 360 €	-159 874 €	2 171 153 €	-53 261 €	7 152 377 €
0,10	26 953	951	4 741	941	6 360 908 €	-271 986 €	1 062 932 €	-37 599 €	7 114 256 €
0,15	29 499	1 309	2 195	583	6 961 764 €	-374 374 €	492 119 €	-23 294 €	7 056 215 €
0,20	30 647	1 569	1 047	323	7 232 692 €	-448 734 €	234 737 €	-12 906 €	7 005 790 €
0,25	31 206	1 708	488	184	7 364 616 €	-488 488 €	109 410 €	-7 352 €	6 978 186 €
0,30	31 477	1 781	217	111	7 428 572 €	-509 366 €	48 651 €	-4 435 €	6 963 422 €
0,35	31 592	1 828	102	64	7 455 712 €	-522 808 €	22 868 €	-2 557 €	6 953 215 €
0,40	31 643	1 860	51	32	7 467 748 €	-531 960 €	11 434 €	-1 279 €	6 945 944 €
0,45	31 677	1 877	17	15	7 475 772 €	-536 822 €	3 811 €	-599 €	6 942 162 €
0,50	31 685	1 885	9	7	7 477 660 €	-539 110 €	2 018 €	-280 €	6 940 288 €

Figure 4.3.7 : Chiffre d'affaire avec les mesures tarifaires

Source : Excel

Comparaison du chiffre d'affaire avant et après mesures

Maintenant que les montants de chiffre d'affaire avec mesure et sans mesure tarifaire ont été déterminés, une comparaison de ces derniers est nécessaire. Sur la figure 4.3.8 ci-dessous, une comparaison des montants de chiffre d'affaire par seuils est déterminée. Lorsqu'aucune mesure n'est appliquée, le taux de chiffre d'affaire chute de 12,5%. Lorsque les mesures sont appliquées, ce taux chute de 9,7% à 12,4% selon le seuil choisi. La deuxième colonne du tableau indique la proportion de chiffre d'affaire que l'application des mesures tarifaires permet de conserver.

	EVOLUTIONS DE CHIFFRE D'AFFAIRE		
Seuils	CA avec mesures / CA sans mesure - 1	CA avec mesures / CA avant résiliations - 1	CA sans mesure / CA avant résiliations - 1
0,05	3,2%	-9,7%	-12,5%
0,06	3,1%	-9,8%	-12,5%
0,10	2,5%	-10,2%	-12,5%
0,15	1,7%	-11,0%	-12,5%
0,20	1,0%	-11,6%	-12,5%
0,25	0,6%	-12,0%	-12,5%
0,30	0,4%	-12,1%	-12,5%
0,35	0,2%	-12,3%	-12,5%
0,40	0,1%	-12,4%	-12,5%
0,45	0,1%	-12,4%	-12,5%
0,50	0,0%	-12,4%	-12,5%

Figure 4.3.8 : Comparaison des évolutions de chiffre d'affaire

Source : Excel

Dans tous les cas, peu importe le seuil choisi, AXA va subir une perte de chiffre d'affaire, avec ou sans application de mesures tarifaires. Néanmoins, une réduction de la prime permet d'augmenter la rétention et de diminuer cette perte de chiffre d'affaire. Ce tableau met en avant le potentiel de rétention de chiffre d'affaire possible en fonction du seuil de binarisation, allant jusqu'à 2,8 points.

Il faudra en complément de cette étude, approfondir les hypothèses retenues avec des analyses spécifiques au périmètre Moto et au comportement des clients de cette branche.

Une approche "Test and learn" peut également être menée. Elle consiste à fixer un seuil de binarisation volontairement élever pour ne sélectionner que les contrats les plus susceptibles de résilier. Les mesures tarifaires sont ensuite appliquées uniquement sur ces individus pour déterminer l'impact réel de la rétention. Six mois plus tard, délai choisi dans le cadre de l'étude, le seuil est remonté ou abaissé selon les résultats réellement obtenus. Cette pratique permet, sur le long terme, de cibler plus précisément les contrats les plus susceptibles de résilier.

Conclusion

L'assurance moto est un marché en plein essor constitué de profils de risques bien spécifiques qu'il est important d'étudier à part entière. De plus, les assurés moto chez AXA sont dans 80% des cas détenteurs de plusieurs contrats au sein de l'entreprise. Un pilotage précis de cette branche est donc nécessaire à chaque moment de la vie du contrat pour conserver les clients en portefeuille. Ce pilotage est notamment nécessaire au moment du remplacement, autrement dit lorsque l'assuré décide de modifier son contrat. La prime d'assurance peut varier de façon significative et entraîner une résiliation. Un modèle de rétention des clients suite à remplacement sur le périmètre moto a donc été mis en place.

Suite à la création et la fiabilisation de la base de données, des analyses univariées ont permis de déterminer plus précisément la structure des remplacements. La sélection des variables à l'aide du calcul des corrélations a permis de réduire le nombre de variables utilisées dans le modèle à 34. La variable cible correspond à une variable binaire qui vaut 1 si la résiliation est consécutive au remplacement et 0 sinon. L'étude du délai entre le remplacement et la résiliation a permis de définir qu'une résiliation est consécutive à un remplacement si cette dernière intervient dans les 180 jours qui suivent ce remplacement. La base totale contient donc 167 929 observations dont 9 616 correspondent à des remplacements résiliés dans les 180 jours.

Afin de caractériser puis prédire cette variable binaire, le modèle linéaire généralisé logit puis la procédure step ont été implémentés sur la base train. Les mesures de qualité de la convergence de ces deux modèles sont assez proches : l'AIC et le BIC sont meilleurs dans le cadre de la procédure step, tandis que la déviance est la même dans les deux cas. Ensuite, les prédictions ont été calculées sur la base test, permettant ainsi de visualiser la performance des modèles. Différents seuils de binarisation des prédictions ont été testés dont le seuil optimal en termes de sensibilité et de spécificité valant 0,06 dans les deux modèles. Là encore, les mesures de performances sont similaires pour les deux modèles et ne permettent pas de trancher. Une validation croisée sur le modèle linéaire généralisé logit a permis de conclure que ce dernier pouvait être généralisé du fait de sa stabilité sur les 10-folds testés. Finalement, les performances des deux modèles étant semblables, la procédure step a été écartée. En effet, le temps d'exécution de cette procédure est bien plus important que celui du GLM et une validation croisée n'a pu être menée sur le step, ne permettant pas de justifier de la généralisation possible du modèle.

L'impact des résiliations suite à remplacement sur le chiffre d'affaire est conséquent. Sans appliquer aucune mesure tarifaire, la perte de chiffre d'affaire générée par les résiliations suite à remplacement est estimée à 12,5% d'après le modèle. Il est supposé que l'application d'une baisse de 5% de la prime permette de conserver 75% des assurés qui résilient pour un motif différent du décès et de la liquidation judiciaire. Dans ce cas, la perte de chiffre d'affaire est limitée à environ 10% et permet la rétention de 2,5% de chiffre d'affaire.

Ces résultats sont néanmoins à prendre avec prudence puisque ce sont des estimateurs théoriques. En effet, le montant de la réduction tarifaire à accorder au client pour éviter une résiliation suite à remplacement a été fixé de manière arbitraire. Il est nécessaire d'effectuer une étude plus approfondie de la sensibilité au prix des individus après leur remplacement pour ajuster la mesure tarifaire et ainsi obtenir une meilleure rétention. Une campagne de Price test est à mettre en place mais l'obtention des résultats est possible seulement au bout de plusieurs années. Cette méthode n'a pas été initiée sur le périmètre moto à ce stade.

Le taux de conservation a lui aussi été déterminé de façon arbitraire, bien qu'il ait été justifié. L'analyse de la sensibilité au prix pour déterminer la meilleure baisse de tarif permettra également de déterminer le taux de conservation ([3] BOUEDDINE, 2013). Une deuxième approche serait de capitaliser sur les résultats obtenus par

le périmètre auto. En effet, l'auto a déjà mis en place une série de price test permettant d'analyser la sensibilité au prix des assurés. Néanmoins, les produits auto et moto sont bien différents et les clients peuvent avoir des comportements différents. La première idée qui consiste à mener une étude sur l'élasticité au prix des assurés moto serait plus judicieuse, bien que coûteuse.

Enfin, le modèle retenu ne permet pas de déterminer un seuil réellement optimal en termes de chiffre d'affaire. Lorsque le taux de conservation passe de 75% à 25% un optimum apparaît pour le seuil 0,15. Il est donc très important de déterminer de façon assez précise un ordre de grandeur du taux de conservation pour optimiser la rentabilité.

Du point de vue de la modélisation, les modèles linéaires généralisés imposent aux observations une loi de probabilité. Cette méthode paramétrique contraint donc la structure même des observations. Pour éviter cette contrainte, il est possible d'avoir recours à des techniques de Data Science comme les Random Forest ou encore les XGBoost. Ces puissants outils de machine learning permettent de prendre en compte la structure même des observations dans la modélisation.

Des améliorations de cette étude sont possibles tant sur l'aspect de la modélisation que sur l'aspect opérationnel. Une autre approche de la modélisation du taux de résiliation à l'aide des modèles de durée peut être envisagée. Cette approche non paramétrique permet de ne pas faire d'hypothèse sur la distribution du taux de résiliation suite à remplacement ([2]BALDE, 2015).

La création d'un outil automatisé qui prendrait en entrée une base de données et permettrait de sortir la perte de chiffre d'affaire potentielle selon le modèle serait intéressante. Le développement d'un tel outil est réalisable à l'aide de R Shiny qui permet de créer une interface aérée avec des résultats qu'il est facile d'exporter.

Maintenant d'un point de vue plus opérationnel : dans le cadre de cette étude et de ce mémoire, les analyses ont été menées uniquement sur le périmètre moto. La perte de chiffre d'affaire dont il est question ne concerne que ce produit. Néanmoins, près de 80% des individus qui ont effectué au moins un remplacement sur leur contrat moto sont multi-détenteurs. La perte de chiffre d'affaire en cas de résiliation pourrait avoir un impact d'autant plus significatif si le client quitte définitivement AXA et résilie l'intégralité de ses contrats, à cause d'une insatisfaction tarifaire au moment de son remplacement moto.

La seule approche technique ne suffira pas. Il faudra également travailler la communication auprès des clients et des distributeurs pour mettre en avant l'avantage accordé et ainsi maximiser l'effet de rétention recherché. Cela pourra par exemple être fait par le biais d'une communication spécifique dans les avis d'échéance.

Références bibliographiques

[1] C. AZENCOTT. *Mettez en place un cadre de validation croisée*, (2020). https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308241-mettez-en-place-un-cadre-de-validation-croisee.

[2] B. BALDE. *Modélisation du taux de résiliation en Assurance MRH*, (2015). Institut des Actuaires.

[3] M. BOUEDDINE. *Assurance Automobile : Analyse de l'impact d'une variation du tarif sur le comportement des assurés lors de l'acte de souscription et de résiliation*, (2013). Institut des Actuaires.

[4] V. COTTET. *Statistique mathématique*, (2019). ENSAE IP Paris.

[5] C. DUTANG. *Actuariat de l'Assurance Non-Vie*, (2020). ENSAE IP Paris.

[6] FFA. *Assurance de biens et de responsabilités - Données clés 2019*, (2020). https://www.ffa-assurance.fr/etudes-et-chiffres-cles/assurances-de-biens-et-de-responsabilite-donnees-cles-par-annee.

[7] ICHI.PRO. *Les courbes ROC et Precision Recall*, (2019). https://ichi.pro/fr/sur-les-courbes-roc-et-precision-recall-213574150750732.

[8] IndexAssurance. *Motifs de résiliation d'une assurance auto*, (2020). https://www.index-assurance.fr/pratique/resiliation/resiliation-d-une-assurance-auto.

[9] A. MELLO. *XGBoost : theory and practice*, (2020). https://towardsdatascience.com/xgboost-theory-and-practice-fb8912930ad6.

[10] ONISR. *Le parc deux-roues motorisés des ménages*, (2020). https://www.onisr.securite-routiere.gouv.fr/etudes-et-recherches/vehicules/parc-des-vehicules/le-parc-deux-roues-motorises-des-menages.

[11] RAKOTOMALALA R. *Analyse de corrélation : Étude des dépendances - Variables quantitatives*, (2017). Université Lumière Lyon 2.

[12] SHAIKH R. *Choosing the right Encoding method-Label vs OneHot Encoder*, (2018). https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b6.

[13] R. RAKOTOMALALA. *Pratique de la Régression Logistique : Régression Logistique Binaire et Polytomique*, (2017). Université Lumière Lyon 2.

[14] R. RAKOTOMALALA. *étude des dépendances - variables qualitatives ; tableau de contingence et mesures d'association*, (2020). Université Lumière Lyon 2.

[15] L. ROUVIERE. *Régression logistique avec R*, (2017). Université Rennes 2.

[16] C. TREMBLAY. *Prédire les sinistres graves en assurance : les apports de l'apprentissage statistique aux modèles linéaires*, (2017). Institut des Actuaires.

Table des figures

Annexes

A - Description détaillée des genres de moto

VEHICULES TOUT TERRAIN

Trial

Moto tout terrain utilisée pour des parcours d'agilité et de franchissement d'obstacles

Cross

Utilisée pour les courses de vitesse sur bosses, non homologuées sur route

Enduro

Utilisée sur les terrains de cross pour la pratique de l'enduro, elle est néanmoins homologuée pour être utilisée sur route.

Quad

Vocation utilitaire même s'il est de plus en plus utilisé comment engin de loisir du fait de sa facilité de conduite et de son accessibilité.

Buggy

Véhicule léger tout terrain de loisir bien qu'utilisable sur route.

Side by Side véhicule (SSV)

Mix entre le quad et le buggy, vocation utilitaire uniquement, équipé d'une benne arrière

VEHICULES ROUTIERS

Basique routière

Moto pour les débutants du fait de sa simplicité de conduite et de son prix abordable.

Trail

Mix entre une moto tout terrain et une moto routière ; utilisée pour une pratique quotidienne

Custom

Utilisée pour de longs trajets du fait de sa faible maniabilité. Réelle personnalisation de la moto par son propriétaire.

Néo Rétro

Véhicule passion doté d'une technologie moderne mais avec un style des 70s

Grand tourisme

Véhicule high tech alliant confort et évasion idéal pour voyager

Trike

Contraction de tricycle et bike, véhicule reflétant un art de vivre et une expérience de liberté. Peu répandu en France.

Roadster

Véhicule sportif et basique en forte évolution ces dernières années

Sportive et hypersportive

Dérivées des motos de compétition, principalement utilisées sur circuits

Scooter 2 & 3 roues

Utilisation quotidienne en zone urbaine du fait sa maniabilité et de sa praticité. Position assise comme une voiture

Supermotard

Moto légère dérivée d'une trail et initialement utilisée comme engin de course

Genres de motos assurés chez AXA

Source : RUN Assurance

B - Code R pour le traitement des données manquantes

```
#######################################    Remplacement des NA    #######################################

# --> varibles quantitatives
num <- base_totale %>% select_if(is.numeric) # sélection l'ensemble des variables numériques
nom_num <- names(num) # permet de récupérer le nom des variables numériques
neg <- -100 # valeur que l'on souhaite attribuer

for(i in 1:ncol(base_totale)){
  if(colnames(base_totale)[i] %in% nom_num){
    replace<-sapply(1:nrow(base_totale),function(k)(if (is.na(base_totale[k,i]) || base_totale[k,i]== ""){
      neg
    }else{
      base_totale[k,i]
    }
    ))
    base_totale[i]<-data.frame(replace)
  }
}
table(is.na(num)) # vérification qu'il n'y a plus de NA

# --> varibles qualitatives
fact <- base_totale %>% select_if(is.factor)
nom_facteurs <- names(fact)

for(i in 1:ncol(base_totale)){
  if(colnames(base_totale)[i] %in% nom_facteurs){
    replace<-sapply(1:nrow(base_totale),function(k)(if (is.na(base_totale[k,i]) || base_totale[k,i]== ""){
      "NR"
    }else{
      base_totale[k,i]
    }
    ))
    base_totale[i]<-data.frame(replace)
  }
}
table(is.na(fact))

# --> varibles temporelles
base_totale <- base_totale %>% mutate(PER_DateNaiss_CP = replace_na(PER_DateNaiss_CP,"1800-01-01"))
# mutate(nom_nouvelle_variable = replace_na(nom_variable_avec_NA, valeur_de_remplacement))
```

Code R pour la modification des NA

Source : Julie Thill

C- Coefficients de corrélation de Pearson

Variable 1	Variable 2	Coefficient de corrélation de Pearson	Valeur absolue du cofficient de corrélation de Pearson	Sens de la corrélation	p value
ecart_cotis	evol_prime_rpl	0,82	0,82	+	-
POL_montantCCA_ap	**taux_cca_ap**	**0,77**	**0,77**	**+**	-
POL_CotisTTC_Tot_ap	nbgar_ap	0,43	0,43	+	-
POL_CotisTTC_Tot_ap	**ecart_cotis**	**0,38**	**0,38**	**+**	-
POL_CotisTTC_Tot_ap	evol_prime_rpl	0,37	0,37	+	-
POL_CT_ap	**anciennete_contrat_au_rpl**	**0,36**	**0,36**	**+**	-
POL_CT_ap	evol_CT_rpl	0,35	0,35	+	-
nbgar_ap	**evol_prime_rpl**	**0,30**	**0,30**	**+**	-
nbgar_ap	ecart_cotis	0,26	0,26	+	-
POL_CotisTTC_Tot_ap	**POL_montantCCA_ap**	**0,25**	**0,25**	**+**	-
POL_montantCCA_ap	nbgar_ap	0,19	0,19	+	-
ANCCLI	**anciennete_contrat_au_rpl**	**0,18**	**0,18**	**+**	-
POL_montantCCA_ap	ecart_cotis	0,13	0,13	+	-
POL_CotisTTC_Tot_ap	**anciennete_contrat_au_rpl**	**0,13**	**0,13**	-	-
POL_montantCCA_ap	evol_prime_rpl	0,12	0,12	+	-
POL_CT_ap	**nb_rpl**	**0,10**	**0,10**	-	-
nbgar_ap	evol_CT_rpl	0,09	0,09	-	-
POL_CotisTTC_Tot_ap	**POL_CT_ap**	**0,09**	**0,09**	**+**	-
POL_CotisTTC_Tot_ap	ANCCLI	0,09	0,09	-	-
POL_CT_ap	**nbgar_ap**	**0,09**	**0,09**	-	**0,00**
nbgar_ap	nb_rpl	0,09	0,09	+	0,00
nb_rpl	**evol_prime_rpl**	**0,09**	**0,09**	**+**	**0,00**
evol_CT_rpl	evol_prime_rpl	0,08	0,08	+	0,00
ecart_cotis	**evol_CT_rpl**	**0,08**	**0,08**	**+**	**0,00**
POL_CotisTTC_Tot_ap	nb_rpl	0,07	0,07	+	0,00
POL_CotisTTC_Tot_ap	**taux_cca_ap**	**0,07**	**0,07**	-	**0,00**
evol_prime_rpl	anciennete_contrat_au_rpl	0,07	0,07	-	0,00
nb_rpl	**taux_cca_ap**	**0,07**	**0,07**	-	**0,00**
evol_CT_rpl	taux_cca_ap	0,06	0,06	+	0,00
ecart_cotis	**anciennete_contrat_au_rpl**	**0,05**	**0,05**	-	**0,00**
POL_CT_ap	evol_prime_rpl	0,05	0,05	-	0,00
POL_montantCCA_ap	**ANCCLI**	**0,05**	**0,05**	-	**0,00**
POL_montantCCA_ap	evol_CT_rpl	0,04	0,04	+	0,00
delai_term_rpl_centre	**evol_prime_rpl**	**0,04**	**0,04**	-	**0,00**
nbgar_ap	taux_cca_ap	0,04	0,04	+	0,00
nb_rpl	**evol_CT_rpl**	**0,04**	**0,04**	**+**	**0,00**
POL_CotisTTC_Tot_ap	delai_term_rpl_centre	0,04	0,04	-	0,00
ecart_cotis	**nb_rpl**	**0,04**	**0,04**	**+**	**0,00**
nb_rpl	anciennete_contrat_au_rpl	0,04	0,04	-	0,00
POL_CotisTTC_Tot_ap	**ELR_TOT_h_prime_ass_ap**	**0,04**	**0,04**	-	**0,00**
nbgar_ap	anciennete_contrat_au_rpl	0,03	0,03	-	0,00
POL_CT_ap	**ecart_cotis**	**0,03**	**0,03**	-	**0,00**
ecart_cotis	delai_term_rpl_centre	0,03	0,03	-	0,00
POL_montantCCA_ap	**nb_rpl**	**0,03**	**0,03**	-	**0,00**
ANCCLI	nb_rpl	0,03	0,03	+	0,00
delai_term_rpl_centre	**anciennete_contrat_au_rpl**	**0,03**	**0,03**	-	**0,00**
evol_prime_rpl	taux_cca_ap	0,03	0,03	-	0,00
POL_CT_ap	**taux_cca_ap**	**0,02**	**0,02**	-	**0,00**
POL_montantCCA_ap	anciennete_contrat_au_rpl	0,02	0,02	-	0,00
taux_cca_ap	**anciennete_contrat_au_rpl**	**0,02**	**0,02**	**+**	**0,00**
nbgar_ap	delai_term_rpl_centre	0,02	0,02	-	0,00
nb_rpl	**delai_term_rpl_centre**	**0,02**	**0,02**	-	**0,00**
ELR_TOT_h_prime_ass_ap	taux_cca_ap	0,02	0,02	+	0,00
POL_CT_ap	**ELR_TOT_h_prime_ass_ap**	**0,02**	**0,02**	-	**0,00**
ELR_TOT_h_prime_ass_ap	nb_rpl	0,02	0,02	+	0,00
ELR_TOT_h_prime_ass_ap	**anciennete_contrat_au_rpl**	**0,02**	**0,02**	**+**	**0,00**
ANCCLI	taux_cca_ap	0,02	0,02	-	0,00
ELR_TOT_h_prime_ass_ap	**evol_prime_rpl**	**0,01**	**0,01**	-	**0,00**
delai_term_rpl_centre	evol_CT_rpl	0,01	0,01	-	0,00
delai_term_rpl_centre	**taux_cca_ap**	**0,01**	**0,01**	**+**	**0,00**
POL_CotisTTC_Tot_ap	evol_CT_rpl	0,01	0,01	-	0,00
ecart_cotis	**ELR_TOT_h_prime_ass_ap**	**0,01**	**0,01**	-	**0,00**
ANCCLI	ecart_cotis	0,01	0,01	+	0,00
ANCCLI	**evol_CT_rpl**	**0,01**	**0,01**	**+**	**0,00**
ANCCLI	ELR_TOT_h_prime_ass_ap	0,01	0,01	+	0,00
ANCCLI	**evol_prime_rpl**	**0,01**	**0,01**	**+**	**0,00**
evol_CT_rpl	anciennete_contrat_au_rpl	0,01	0,01	+	0,00
ELR_TOT_h_prime_ass_ap	**evol_CT_rpl**	**0,01**	**0,01**	-	**0,00**
POL_montantCCA_ap	POL_CT_ap	0,01	0,01	+	0,00
POL_montantCCA_ap	**ELR_TOT_h_prime_ass_ap**	**0,00**	**0,00**	**+**	**0,09**
POL_CT_ap	delai_term_rpl_centre	0,00	0,00	-	0,17
nbgar_ap	**ELR_TOT_h_prime_ass_ap**	**0,00**	**0,00**	**-**	**0,30**
ecart_cotis	taux_cca_ap	0,00	0,00	+	0,47
POL_CT_ap	**ANCCLI**	**0,00**	**0,00**	-	**0,51**
ELR_TOT_h_prime_ass_ap	delai_term_rpl_centre	0,00	0,00	+	0,62
nbgar_ap	**ANCCLI**	**0,00**	**0,00**	**+**	**0,65**
ANCCLI	delai_term_rpl_centre	0,00	0,00	-	0,74
POL_montantCCA_ap	**delai_term_rpl_centre**	**0,00**	**0,00**	**+**	**0,95**

Coefficients de corrélation de Pearson

Source : Sortie R et mise en page Excel