

INSEA

Projet de Fin d'Etudes

Etude de méthodes de tarification sur un portefeuille automobile

Préparé par : *Mme. Chaymaa CHICHANE*
M. Mohamed Reda JOULID

Sous la direction de : *M. Khalid ZOUHAR (INSEA)*
M. Adil BENSOUNA (SANLAM)

Soutenu publiquement comme exigence partielle en vue de l'obtention du

Diplôme d'Ingénieur d'Etat

Filière : Actuariat-Finance

Devant le jury composé de :

- *M. ZOUHAR KHALID (INSEA)*
- *M. MARRI FOUAD (INSEA)*
- *M. BENSOUNA ADIL (SANLAM)*
- *M. MRICHA OMAR (SANLAM)*

Dédicace

Nous dédions cette œuvre

À nos chers parents qui sont toujours présents pour leurs soutiens, Leur dévouement et leur investissement énergétique dans la création d'un environnement propice à nos études ont été inestimables. Leur soutien indéfectible et leur encouragement ont été les piliers de notre réussite.leur amour a fait de nous ce que nous sommes aujourd'hui. Que Dieu les garde et les protège,

À nos frères et sœurs en signe d'amour, de reconnaissance et de gratitude pour le dévouement et les sacrifices dont ils ont fait toujours preuve à notre égard,

À nos chers professeurs qui ont suscité notre admiration pour leur savoir, leur efficacité et leur compétence,

À tous ceux qui nous aiment.

Remerciements

En premier lieu, nous tenons à remercier notre encadrant de stage, Monsieur BENSOUNA ADIL. Puisqu'il a su nous faire confiance lors de cette aventure et a partagé ses connaissances de manière très pédagogique. Nous le remercions aussi pour sa disponibilité et la qualité de son encadrement,

Nous tenons à exprimer notre profonde gratitude envers M. MRICHA OMAR pour sa disponibilité constante tout au long du stage, ainsi que pour ses efforts inlassables et son soutien précieux. Un remerciement particulier est également adressé à M. EL MALEKY ZAKARIA pour son accueil chaleureux et sa disponibilité sans faille.

Nous souhaitons également exprimer notre sincère reconnaissance à notre encadrant académique, M. KHALID ZOUHAR, pour ses conseils éclairés, ses remarques pertinentes et son implication inestimable dans ce travail.

Enfin, nous tenons à exprimer notre gratitude à toutes les personnes qui ont contribué, directement ou indirectement, à l'élaboration de ce travail.

Résumé

Les avancées fulgurantes dans le domaine de la science des données, en particulier dans le domaine de l'apprentissage statistique, ont révolutionné les méthodes actuarielles, notamment en ce qui concerne la tarification. Dans ce mémoire, nous nous intéressons à l'étude de méthodes de tarification sur un portefeuille automobile. Pour ce faire, notre rapport sera divisé en six chapitres:

- Le premier concerne la présentation du contexte d'étude
- Le deuxième porte sur le traitement et l'analyse des données
- Le troisième traite la construction d'un nouveau zonier tarifaire
- Le quatrième aborde la tarification par GLM
- Le cinquième porte sur la tarification par les méthodes d'apprentissage
- Le dernier examine la comparaison des différents modèles

Mots clés

Assurance automobile
Tarification
GLM
apprentissage automatique
prime pure
CART
XGBOOST

Table des matières

Dédicace	3
Remerciements	4
Résumé	5
Mots clés	5
Table des figures	10
Introduction générale	12
Chapitre 1 : Environnement et contexte de l'étude	13
I. Introduction	14
II.Présentation du contexte d'étude	14
II.1.Aperçu concis sur l'organisme d'accueil	14
II.2.L'assurance automobile	17
II.3.Tarif de l'assurance obligatoire "responsabilité civile automobile "	18
II.4.Cadre légale de la tarification au Maroc	18
II.5.Le marché de l'assurance au Maroc	19
III.Modélisation tarifaire en assurance automobile	20
III.1.Principes de la segmentation et la mutualisation	20
III.2.Prime pure et modèle fréquence-cout moyen	21
VI.Conclusion	23
Chapitre 2 : Traitement et analyse des données	24
I. Introduction	25
II.Analyse de la base de données	25
II.1.Description des variables	25
II.2.Epuration et fiabilisation des données	26
II.3.Analyse graphique	26
II.4. Analyse de la corrélation	29
III.Ecrêtement des sinistres graves	30

IV.Conclusion	33
Chapitre 3 : Construction d'un nouveau zonier tarifaire	34
I. Introduction	35
II.Méthodologie de Travail	35
III.Cadre théorique de l'algorithme CART	36
III.1.Principe de construction de l'arbre	36
III.2.l'élagage de l'arbre	38
III.3. Validation croisée	39
IV.Analyse du zonier actuel	40
IV.1.Etude des indicateurs de sinistralité	41
V.Description de la base de données externe	43
V.1.description des variables	43
V.2.Etude des corrélations	44
VI.Application	45
VI.1.Construction du zonier du cout moyen	46
VI.1.1.construction de l'arbre maximal	46
VI.1.2 élagage de l'arbre	46
VI.2. Construction du zonier de la fréquence	50
VII.Conclusion	51
Chapitre 4: Tarification par l'approche classique GLM	52
I.Introduction	53
II.Cadre théorique des modèles linéaires généralisés	53
II.1.La composante aléatoire	54
II.2 La composante déterministe	54
II.3.La fonction de lien	55
II.4.Sélection des variables explicatives	55
II.5.Significativité des variables	56
II.6.Evaluation de la qualité d'ajustement	56
II.7.Choix du meilleur modèle	57

III.Application	57
III.1.segmentation des variables	57
III.2.Modélisation de la fréquence	58
III.2.1 Analyse des modèles candidats	58
III.2.2.mise en oeuvre des modèles	59
III.2.3.Analyse graphique des résidus	64
III.3. Modélisation du coût moyen	65
III.3.1. Analyse des modèles candidats	65
III.3.2.mise en oeuvre des modèles	65
III.3.3.Analyse graphique des résidus	67
IV.Conclusion	68
Chapitre 5: Tarification par des méthodes d'apprentissage	69
I.Introduction	70
II.Cadre théorique de l'algorithme XGBOOST	70
III.Application de l'algorithme XGBOOST	72
III.1. Modélisation de la fréquence	73
III.1.1.Recherche du meilleur nombre d'itérations	73
III.1.2.Importance des variables	74
III.2.Modélisation du coût moyen	75
III.2.1.Recherche du meilleur nombre d'itérations	75
III.2.2.Importance des variables	75
IV.Application de l'algorithme CART	76
IV.1.Modélisation de la fréquence	76
IV.1.1.Arbre maximal	76
IV.1.2.Elagage de l'arbre	77
IV.2.Modélisation du coût moyen	78
IV.2.1.Arbre maximal	78
IV.2.2.Elagage de l'arbre	79
V.Conclusion	80
Chapitre 6: Comparaison des différents modèles	81
I.Introduction	82
II.Comparaison de l'Erreur Quadratique Moyenne	82
III.Synthèse des primes pures	84

Table des matières

IV.construction d'une application VBA Excel	88
V.Conclusion	88
Conclusion générale	90
Bibliographie et Webographie	91
ANNEXES	93

Table des figures

1	historique	15
2	Evolution des Chiffres Clés	15
3	Evolution du chiffre d'affaires par branches	16
4	Evolution des primes acquises pour l'activité non vie	17
5	Le tarif de l'usage "A"	18
6	les primes emises pour l'exercice de 2023	19
7	Principes de la segmentation	20
8	les éléments de la prime commerciale	22
9	Distribution du nombre de sinistres	27
10	Distribution du cout moyen	27
11	Coefficient de Pearson	29
12	Coefficient V de Cramer	30
13	FME pour le segment matériel	32
14	FME pour le segment corporel	32
15	Illustration de l'arbre CART	37
16	Séparation des données pour la méthode validation croisée du 5-fold	40
17	Carte du risque actuel projetée avec SAS	40
18	la Fréquence des sinistres	41
19	le coût moyen des sinistres	41
20	ratio de sinistralité	42
21	matrice de corrélation	44
22	L'arbre de décision maximal pour le coût moyen	46
23	Graphique de l'erreur de la validation croisée	47
24	la sortie cptable de l'arbre maximal	47
25	l'arbre optimal pour le coût moyen	48
26	l'importance des variables dans le modèle du cout moyen	49
27	l'arbre optimal pour la fréquence	50
28	l'importance des variables dans le modèle de la fréquence	51
29	Lois Utiles de la Famille Exponentielle	55
30	La segmentation selon la fréquence des sinistres	58
31	La segmentation selon le coût moyen	58
32	Ajustement de la fréquence des sinistres par la loi de poisson et la loi Binomiale Négative	59
33	Les variables sélectionnées par l'approche Forward	60
34	Estimation des paramètres du modèle Binomial Négatif	60
35	Evaluation de la qualité d'ajustement du GLM pour la loi Binomiale Négative	61
36	Estimation des paramètres du modèle de poisson	61
37	Evaluation de la qualité d'ajustement du GLM pour la loi de poisson	62
38	Estimation des paramètres de la loi ZIP	62

39	Evaluation de la qualité d'ajustement du GLM pour la loi ZIP . . .	63
40	Estimation des paramètres de la loi ZINB	63
41	Evaluation de la qualité d'ajustement du GLM pour la loi ZINB	64
42	analyse des résidus pour le modèle ZINB	64
43	QQplot des lois Gamma et Log-Normale	65
44	sélection des variables par l'approche Forward	66
45	Evaluation de la qualité d'ajustement du GLM pour la loi gamma et log-normale	66
46	L'estimation des paramètres pour la loi log-normale	67
47	Analyse graphique des résidus pour le modèle Log-Normal	67
48	illustration du fonctionnement de XGBOOST	72
49	Evolution de la RMSE sur la base d'apprentissage en fonction du nombre d'itérations	73
50	Importance des variables dans la modélisation de la fréquence . .	74
51	Evolution de la RMSE sur la base d'apprentissage en fonction du nombre d'itérations	75
52	Importance des variables dans la modélisation du coût moyen . .	76
53	l'arbre maximal de la fréquence	77
54	Graphique de l'erreur de la validation croisée	77
55	La sortie cptable de l'arbre maximal	78
56	L'arbre optimal de la fréquence	78
57	L'arbre maximal du coût moyen	79
58	Graphique de l'erreur de la validation croisée	79
59	La sortie cptable de l'arbre maximal	80
60	l'arbre optimal du coût moyen	80
61	Distribution des primes pures pour le modèle GLM1	84
62	Distribution des primes pures pour le modèle GLM2	85
63	Distribution des primes pures pour le modèle CART	86
64	Distribution des primes pures pour le modèle XG-Boost	87
65	interface de l'application VBA	88
66	Distribution de combustion	93
67	Distribution de la puissance fiscale	93
68	Distribution de CRM	93
69	Distribution de sexe	93
70	Distribution de l'âge du conducteur	94
71	Distribution de l'âge du véhicule	94
72	Distribution des régions	94
73	Distribution de la situation matrimoniale	94
74	segmentation proposée par l'organisme	97

Introduction générale

Les acteurs du secteur de l'assurance sont confrontés à un monde en perpétuelle mutation, caractérisé par des défis technologiques et des évolutions dans les comportements et les pratiques. Parallèlement, ils doivent s'adapter à un environnement de taux bas persistant et continuer à se transformer pour répondre à des exigences réglementaires de plus en plus strictes.

Ces dernières années, la quantité d'informations collectées auprès des assurés a considérablement augmenté. De nombreuses recherches et efforts ont été entrepris pour améliorer le modèle classique GLM utilisé par les actuaires pour la tarification.

En outre, les avancées récentes en science des données ont permis l'émergence de nouvelles idées et d'applications. L'utilisation de l'apprentissage statistique, notamment, présente de nouvelles possibilités pour l'analyse des données et la création de modèles. Dans différents domaines (web et multimédia, trading, biologie, assurances, etc.), permettant de transformer les données collectées en informations fiables et d'établir des prédictions plus précises pour éclairer les prises de décision.

Dans ce contexte, l'objectif de notre étude est de démontrer l'apport des algorithmes d'apprentissage dans la tarification de l'assurance RC-automobile et précisément pour la responsabilité civile

Chapitre 1 : Environnement et contexte de l'étude

I. Introduction

L'assurance est une opération qui permet à une personne ou à une organisation de se prémunir contre les conséquences financières de la survenance d'un événement non souhaitable que l'on appelle risque, en effet l'assuré souscrit un contrat d'assurance couvrant des risques pouvant être de différentes natures (accident de voitures, maladie, vol, . . .), le souscripteur verse une prime (ou cotisation) à l'assureur, l'assureur promet en contre partie dans le cadre du contrat d'assurance la réparation du préjudice survenu ou le service d'une prestation. L'assurance automobile, objet du présent mémoire, est l'une des principales composantes de l'assurance IARD (Incendies, Accidents et Risques Divers) qui permet de protéger les biens. Ce chapitre se propose de détailler le contexte de notre étude. Nous commencerons par présenter l'organisme d'accueil, puis nous examinerons le marché de l'assurance au Maroc, avant d'aborder les principes fondamentaux de la tarification en non vie.

II. Présentation du contexte d'étude

II.1. Aperçu concis sur l'organisme d'accueil

Sanlam est un groupe financier de référence qui est coté sur le JSE Limited (Bourse de Johannesburg) et sur le Namibian Stock Exchange (Bourse de la Namibie). Fondé en 1918 en tant que compagnie d'assurance vie, le groupe Sanlam, basé en Afrique du Sud, s'est par la suite transformé en une entreprise de services financiers diversifiés. Présente sur le marché des assurances au Maroc depuis 1949, Sanlam Maroc joue un rôle de premier plan dans les secteurs de l'assurance vie et non-vie. Elle occupe une position de leader dans le domaine de l'assurance Non-Vie, se classant au premier rang pour les assurances automobile et santé. Avec un réseau de plus de 481 agents généraux, Sanlam bénéficie du réseau exclusif le plus vaste au Maroc. Cette ampleur lui permet d'assurer une présence régionale significative et de mettre en œuvre une politique de proximité efficace avec l'ensemble de sa clientèle.

La figure suivante récapitule les différentes dates clés qui ont marqué l'histoire de l'organisme :

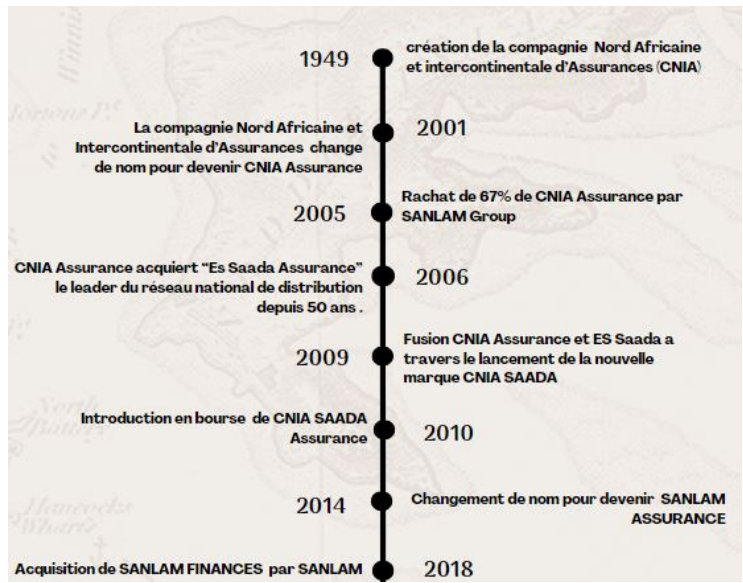


Figure 1: historique

- Evolution des Chiffres Clés :

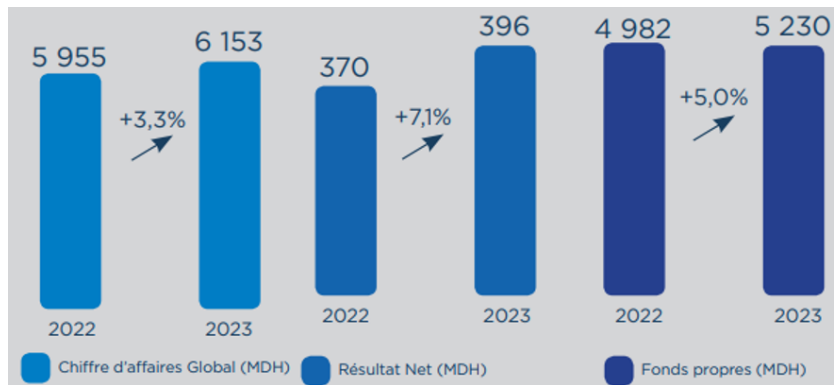


Figure 2: Evolution des Chiffres Clés

La croissance significative du chiffre d'affaires global de SANLAM en décembre 2023, comme le montrent les graphiques ci-dessus, est principalement attribuable à une augmentation de 6,4% dans l'activité Non-Vie, qui a totalisé 5 346 MDH de chiffre d'affaires Non-vie à la fin de 2023 par rapport à l'année précédente. Cette expansion est également accompagnée d'une augmentation

remarquable du résultat net de la compagnie, enregistrant une hausse de 7,1%, pour atteindre 396 MDH à la fin de l'année 2023 par rapport à 2022.

Parallèlement, les fonds propres de la compagnie ont également connu une progression de 5%, passant de 4 982 millions de dirhams à la fin de 2022 à 5 230 MDH à la fin de 2023, démontrant ainsi une performance financière solide et soutenable.

- Evolution du chiffre d'affaires par branches :

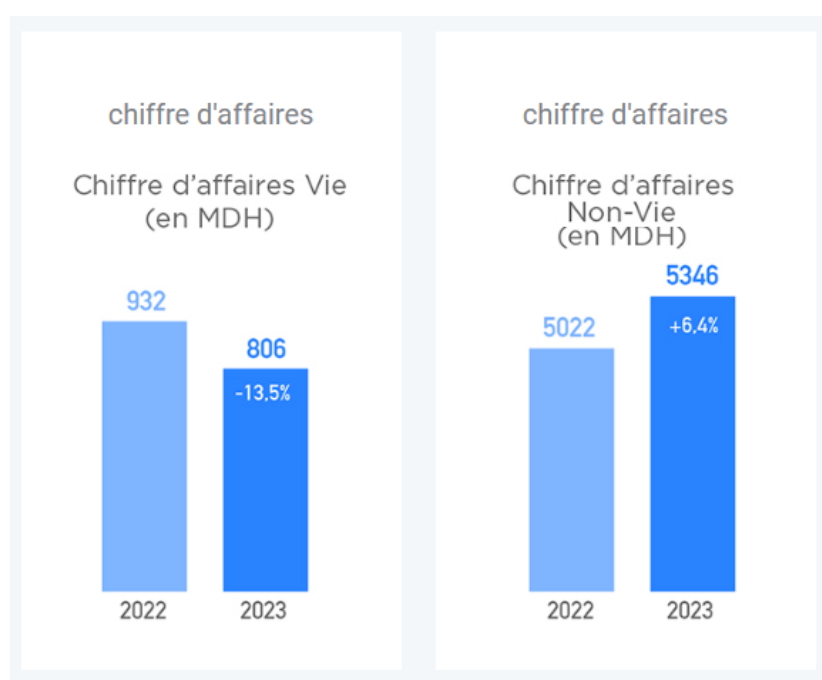


Figure 3: Evolution du chiffre d'affaires par branches

Le chiffre d'affaires de l'activité vie de Sanlam a connu une diminution de 13,5% entre 2022 et 2023, passant de 932 MDH à 806 MDH, tandis que celui de la branche non vie a augmenté de 6,4%, passant de 5022 MDH à 5346 MDH sur la même période. La branche non vie représente une part significative du chiffre d'affaires global de Sanlam, contribuant ainsi de manière substantielle à sa performance financière globale.

- Evolution des primes acquises pour l'activité non vie :

En millions de dirhams

	1er Sem. 2021	1er Sem. 2022	1er Sem. 2023	Evolution S12023/S12022	Part marché
Sanlam	2 611,7	2 841,3	3 075,4	8,2%	17,1%
Wafa Assurance	2 456,7	2 577,3	2 947,3	14,4%	16,4%
AtlantaSanad	2 313,5	2 391,8	2 478,1	3,6%	13,8%
Axa Assurance Maroc	1 982,8	2 123,1	2 274,9	7,1%	12,7%
RMA	2 069,8	2 119,3	2 248,8	6,1%	12,5%
MAMA	860,0	924,3	965,8	4,5%	5,4%
Allianz	652,3	646,2	711,9	10,2%	4,0%
CAT	668,1	713,2	711,1	-0,3%	4,0%
MAMDA	501,5	516,4	621,2	20,3%	3,5%
MATU	385,0	463,9	570,6	23,0%	3,2%

Figure 4: Evolution des primes acquises pour l'activité non vie

Le classement national des compagnies d'assurance marocaines dans le secteur de l'assurance Non-Vie met en évidence la constante ascension des primes émises par SANLAM, confirmant ainsi son leadership dans ce domaine. Au cours des trois dernières années, les primes émises par SANLAM ont enregistré une progression significative, passant de 2 611,7 MDH en 2021 à 3 075 MDH en 2023. Cette évolution positive représente une augmentation de 8,2 % entre le premier semestre de 2023 et celui de 2022, consolidant ainsi sa position avec la plus grande part de marché, qui s'élève à 17,1 %.

II.2. L'assurance automobile

L'assurance automobile est un type d'assurance qui couvre les risques associés à la conduite d'un véhicule, offrant une protection financière en cas d'accidents pour le conducteur et les tiers. Elle se divise généralement en deux parties : l'assurance obligatoire et l'assurance facultative.

- Responsabilité civile :

La première partie du contrat d'assurance auto concerne la garantie de responsabilité civile qui est une garantie obligatoire pour tout véhicule motorisé, se divisant en deux types : la responsabilité civile matérielle, couvrant les dommages aux biens d'autrui, et la responsabilité civile corporelle, prenant en charge les dommages physiques causés à une tierce personne. Ces garanties sont essentielles pour assurer la protection des tiers en cas d'accident. Un automobiliste qui n'a pas souscrit cette garantie risque des amendes, la suspension de son permis ou même la mise en fourrière lors d'un contrôle.

- Garanties annexes :

La deuxième partie de l'assurance automobile concerne les garanties facultatives. Celles-ci incluent la couverture contre les collisions, le vol, l'incendie du véhicule, le bris de glace, la protection des accessoires embarqués dans la voiture.

II.3. Tarif de l'assurance obligatoire "responsabilité civile automobile "

Le tarif de l'assurance " Responsabilité civile " est en fonction de l'usage et des caractéristiques du véhicule :

1. Puissance fiscale (force en chevaux fiscaux) et cylindrée (en cm³) pour les véhicules de moins de 3,5 tonnes ;
2. Type du moteur (essence ou diesel) pour les véhicules de moins de 3,5 tonnes ;
3. Nombre de places pour les véhicules relevant de l'usage "B" ;
4. Poids total autorisé en charge.

Dans notre rapport, nous nous intéressons à l'usage "A", qui est celui des véhicules utilisés pour les besoins privés ou pour les besoins d'une profession ou d'une activité à l'exclusion des véhicules utilisés pour le commerce. Son tarif est fixé comme suit :

Puissance Fiscale		Prime (en DH)
Moteur à Essence	Moteur de Type Diesel	
Jusqu'à 6 cv	Jusqu'à 4 cv	1840
7 et 8 cv	5 cv	2238
9 et 10 cv	6 et 7 cv	2429
11 cv et plus	8 cv et plus	3490

Figure 5: Le tarif de l'usage "A"

II.4. Cadre légale de la tarification au Maroc

En 2006, le Maroc a ouvert la voie à la déréglementation de l'assurance responsabilité civile automobile, offrant ainsi une plus grande souplesse aux assureurs tout en introduisant des règles tarifaires. Malgré cette libéralisation, les tarifs sont restés stables depuis lors, ce qui a entraîné une concurrence acharnée entre les compagnies d'assurance pour se différencier par des services et garanties supplémentaires. La circulaire de l'ACAPS de 2016 est cruciale dans ce contexte car elle précise les critères pour calculer les primes d'assurance, incluant l'usage

du véhicule, sa puissance fiscale, etc.

Parallèlement, la législation marocaine s'est adaptée pour mieux encadrer le secteur des assurances. La loi n° 110-14, adoptée dans cet objectif, instaure un régime de couverture des événements catastrophiques, modifiant ainsi la loi précédente (n° 17-99) pour définir les responsabilités des compagnies d'assurance et les droits des assurés. Cette loi rend obligatoire l'assurance contre les événements catastrophiques pour certains contrats, couvrant une gamme étendue de risques allant des catastrophes naturelles aux actions humaines violentes comme le terrorisme.

II.5. Le marché de l'assurance au Maroc

Primes émises (*)

	Primes (2023)	Evolution (vs n-1)
Vie	25 852,5	1,8%
Epargne-Support Dirhams	21 262,9	1,5%
Décès	3 332,9	3,7%
Epargne-Support Unités de Compte	1 256,2	2,0%
Autres opérations	0,5	218,0%
Non vie	30 074,2	5,8%
Evènements catastrophiques	571,2	1,9%
Accidents corporels	5 389,6	6,8%
dont maladie	4 602,0	5,7%
AT & MP	2 557,0	2,5%
Automobile	14 370,5	4,7%
dont RC	11 919,9	5,4%
RC Générale	741,5	5,4%
Incendie	2 278,1	8,8%
Risques techniques	459,2	61,4%
Transport	839,9	-4,9%
Assistance	1 483,1	7,1%
Crédit - caution	288,0	-1,2%
Autres opérations	1 096,2	15,4%
Acceptations	765,4	11,6%
Vie	-	-
Non vie	765,4	11,6%
Total	56 692,0	4,0%

Figure 6: les primes émises pour l'exercice de 2023

Au Maroc, les primes, à l'exception des réassureurs exclusifs, ont progressé de 4% à 56 692 millions de dirhams en 2023, comme indiqué par l'Autorité de contrôle des assurances et de la prévoyance sociale (Acaps) du Royaume. La croissance la plus forte a été observée au niveau des primes des sociétés d'assurance non-vie. Ces primes ont atteint 30 074,2 millions de dirhams, une hausse de 5,8% en glissement annuel. Les primes d'assurance automobile demeurent les plus élevées du secteur non-vie. Elle se situe à 14 370,5 milliards de dirhams, une hausse de 4,7%.

En ce qui concerne le secteur vie, les primes dans cette catégorie ont atteint

25 850,5 millions de dirhams, une hausse de 1,8% Elles ont été soutenues par les polices épargne supports-dirhams et décès, qui ont enregistré des primes en hausse de 1,5% pour le premier et 3,7% pour le second.

III.Modélisation tarifaire en assurance automobile

La tarification des produits d'assurance automobile est un processus complexe et crucial pour les compagnies d'assurance. Elle implique une analyse approfondie de nombreux critères liés au véhicule et à son conducteur, tels que l'âge, le type de véhicule, l'historique de conduite, etc. Cette tarification repose sur une combinaison d'analyses rétrospectives et prospectives, tout en tenant compte des objectifs de rentabilité et de la concurrence sur le marché de l'assurance.

III.1.Principes de la segmentation et la mutualisation

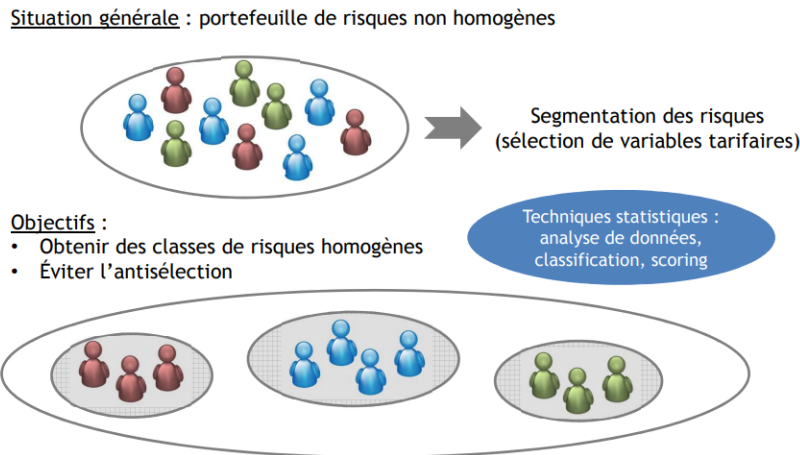


Figure 7: Principes de la segmentation

La segmentation tarifaire en assurance découle de la diversité des risques au sein des portefeuilles d'assurance. Si les risques étaient homogènes, il n'y aurait pas de nécessité d'appliquer des tarifs différents à chaque risque, et la prime d'assurance serait uniforme pour tous les assurés. La segmentation, en assurance, consiste à classer les risques en fonction de différents critères afin d'ajuster au mieux la prime en fonction de l'importance du risque que représente chaque assuré. Cela permet à l'assureur de déterminer les conditions de son engagement.

Deux principes se présentent comme contradictoires en assurance : la segmentation et la mutualisation. La mutualisation des risques est nécessaire pour faire face à la grande variabilité des risques et réduire l'exposition globale au risque. Elle repose sur la loi des grands nombres et suppose que les dommages subis par un groupe d'individus sont des variables aléatoires indépendantes et identiquement distribuées. En revanche, le principe de segmentation consiste à regrouper les risques similaires pour appliquer des primes différenciées à chaque groupe. Cela permet de distinguer les bons risques des mauvais risques en découpant le portefeuille d'assurés en sous-groupes, appelés classes de risques, en fonction de caractéristiques distinctives.

Dans un marché très concurrentiel, la segmentation devient incontournable pour les assureurs. Elle leur permet de fidéliser leur portefeuille et d'attirer de nouveaux clients représentant les bons risques afin d'éviter l'anti-sélection. L'anti-sélection se produit lorsque les assureurs proposent la même prime pour tous les types de risques, ce qui conduit les bons risques à choisir l'offre concurrente moins chère et les mauvais risques à choisir l'offre de l'assureur en question malgré un risque plus élevé.

La limite de la segmentation réside dans la connaissance des risques du portefeuille, nécessaire pour que la loi des grands nombres et le théorème central limite puissent s'appliquer. En conditions réelles, cette connaissance des risques est imparfaite, et la segmentation peut même diminuer cette connaissance et augmenter l'incertitude d'estimation. Cependant, tant que le tarif est construit sur une base de méthodes statistiques, il est nécessaire de disposer d'un volume de données suffisant pour que la loi des grands nombres s'applique et que la volatilité de l'estimateur soit contrôlée.

III.2.Prime pure et modèle fréquence-cout moyen

- La prime pure :

La prime pure est le montant moyen du sinistre que l'assureur s'attend à devoir prendre en charge sur une période donnée, représentant ainsi le coût du risque couvert par le contrat d'assurance.

Pour obtenir la prime commerciale réellement versée par l'assuré, il est nécessaire d'ajouter à la prime pure plusieurs éléments. Tout d'abord, les frais de gestion et d'acquisition, destinés à couvrir les coûts supportés par l'assureur tout au long de la vie du contrat, tels que les frais d'acquisition de nouveaux clients et les frais administratifs liés à la gestion des contrats en portefeuille. Ensuite, il convient d'inclure les frais de sécurité, qui peuvent être proportionnels à la variance du sinistre ou à la prime pure, afin de couvrir le risque de mauvaise tarification.

Une fois ces composantes prises en compte, on peut alors établir l'équation suivante

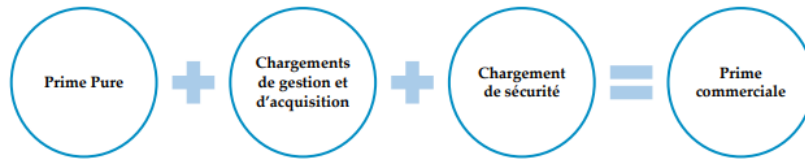


Figure 8: les éléments de la prime commerciale

- Modèle fréquence –coût moyen :

Dans le cadre d'un modèle collectif, la charge financière totale pour la période en question est :

$$S = \sum_{i=1}^N X_i$$

Avec : N: variable aléatoire discrète représentant le nombre de sinistres
 Xi : variable aléatoire continue représentant le montant du ième sinistre

$$(i = 1, \dots, N)$$

Pour simplifier le calcul de la prime pure, nous considérons que le nombre d'événements est indépendant des paiements. Et que les coûts des sinistres ont le même comportement aléatoire On obtient donc :

$$E(S) = E(N) \times E(X)$$

Cette formule permet d'établir une prime pure en estimant séparément la moyenne des fréquences et la moyenne des coûts de sinistre.

En univers segmenté, la relation précédente peut se réécrire

$$E(S/Z) = E(N/Z) \times E(X/Z)$$

où

$$\mathbf{Z} = (Z_1, Z_2, \dots, Z_p)$$

le vecteur des p critères tarifaires retenus.

Une approche intuitive pour déterminer la prime pure consiste à modéliser séparément la fréquence et le coût moyen, puis à multiplier les échelles de risque modélisées précédemment pour obtenir les échelles de prime pure.

L'actuaire dispose de plusieurs types de données pour modéliser la fréquence et le coût moyen d'une garantie donnée, ce qui permet de déterminer sa prime pure. Cependant, il est crucial de souligner que la tarification ne doit jamais reposer sur des informations ex post, c'est-à-dire, des informations connues après le début de l'exposition ou après la déclaration du sinistre. Bien que ces informations puissent être très informatives, elles ne peuvent être utilisées dans le processus de tarification, en effet seuls les critères disponibles au moment de l'établissement du tarif peuvent être exploités pour calculer une prime.

IV. Conclusion

Dans ce chapitre, nous avons exposé brièvement l'organisme d'accueil, en expliquant sa position dans le domaine de l'assurance. Par la suite, nous avons analysé l'évolution des chiffres clés, en mettant en lumière la croissance et les performances récentes de l'organisme, soulignant ainsi sa dynamique de développement et son impact sur le marché.

Nous avons aussi examiné les principes de la segmentation des risques et de la mutualisation, en donnant une explication sur l'application de ces concepts essentiels pour le calcul des primes.

Enfin, nous avons présenté les principes de la tarification en assurance non-vie, en mettant l'accent sur les méthodes et les modèles utilisés pour évaluer les risques et établir les primes.

Chapitre 2 : Traitement et analyse des données

I. Introduction

Dans cette section, nous allons présenter la base de données ainsi que les différentes variables et traitements effectués. Avant de débiter les travaux de modélisation, cette étape revêt une importance capitale, car elle apporte des renseignements essentiels sur les résultats prévus et leur qualité.

Dans notre situation, cette approche d'analyse nous donnera l'occasion de réaliser une première évaluation de la qualité et de la fiabilité des informations. Par ailleurs, cette étude expliquera notre démarche qui consiste à étudier de manière distincte les deux parties de la branche RC, la responsabilité civile corporelle et matérielle.

II. Analyse de la base de données

II.1. Description des variables

Notre analyse porte sur la garantie RC automobile, Les données utilisées dans notre étude Couvrent les quatre exercices de 2019 à 2022. Elles sont réparties en deux tables : table "production" et table "sinistres".

La table "production" englobe des informations sur le contrat d'assurance, le type de véhicule et les détails de l'assuré. Ces données sont enregistrées à travers différentes variables qui contribuent à définir le profil de risque de chaque contrat.

Variable	Description
NUMERO_CONTRAT	numéro de contrat
DEBUT	date de début du contrat
ANNEE_EFFECT	année d'effet du contrat
FIN	date de fin du contrat
COMBUSTION	Le type de combustible du véhicule
PUISSANCE_FISCALE	puissance fiscale du véhicule
ANCIENNETE_MEC	date de mise en circulation du véhicule
CRM	Coefficient de Réduction/Majoration
SEXE	Sexe du conducteur
DATE_NAISSANCE	Date de naissance du conducteur
SITUATION_MATRIMONIALE	situation matrimoniale du conducteur
REGION	Région
ZONE_PDV	Le niveau de risque de la zone
PROVINCE	Province
EXPOSITION	l'exposition au risque

Table 1: Description des variables de la base production

La table "sinistres" englobe les données relatives aux sinistres :

la Variable	Description
Numero_Sinistre	Numéro de Sinistre
Numero_contrat	numéro de contrat
Date_ouverture	Date d'ouverture du sinistre
Date_survenance	Date de survenance du sinistre
Description_Garantie	Description de la garantie incluse dans le contrat
Date_Clature	Date de clôture du sinistre
Charge_avec_Recours	charge du sinistre incluant les recours

Table 2: Description des variables de la base sinistre

II.2.Epuration et fiabilisation des données

Le processus d'épurement et de traitement des bases de données, est nécessaire avant toute mise en œuvre des résultats.

Cette étape est cruciale pour assurer la fiabilité des données utilisées et éviter toute distorsion susceptible de biaiser nos résultats.

Variable	Nombre de valeurs manquantes	Pourcentage
SEXE	1270	0.03%
ANCIENNETE_MEC	531	0.01%
CRM	589	0.01%

Table 3: les valeurs manquantes de la base des données

Les différents traitements effectués sont décrits comme suit :

- Ajustement des variables pour correspondre à leur format approprié, incluant la conversion en format numérique, format de chaîne de caractères et format temporel
- Élimination des valeurs manquantes, qu'il s'agisse d'espaces ou de points (en vue de leur faible nombre)
- Calcul des âges du conducteur et du véhicule
- Correction des Valeurs Aberrantes et Normalisation des Données.

II.3.Analyse graphique

dans cette partie , nous allons Analyser l'évolution de la fréquence et le coût en fonction des variables tarifaires

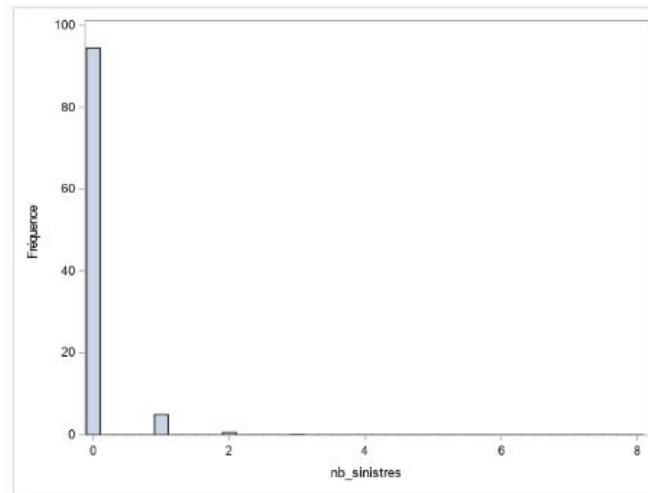


Figure 9: Distribution du nombre de sinistres

L'histogramme du nombre de sinistres montre un pic à zéro, indiquant que la majorité des assurés n'ont pas déclaré de sinistres pendant la période d'observation. Cette observation influence significativement la moyenne du nombre de sinistres.

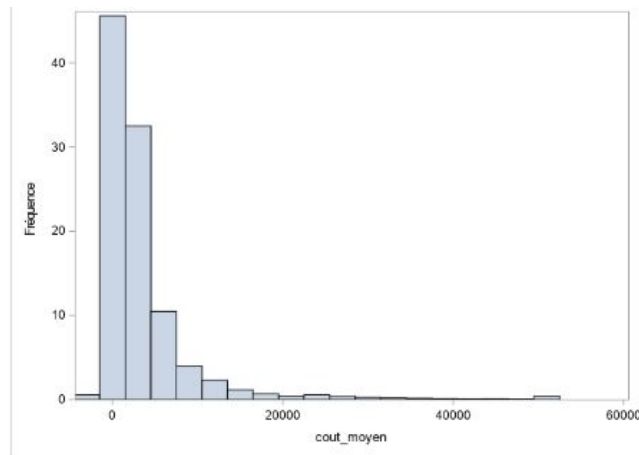
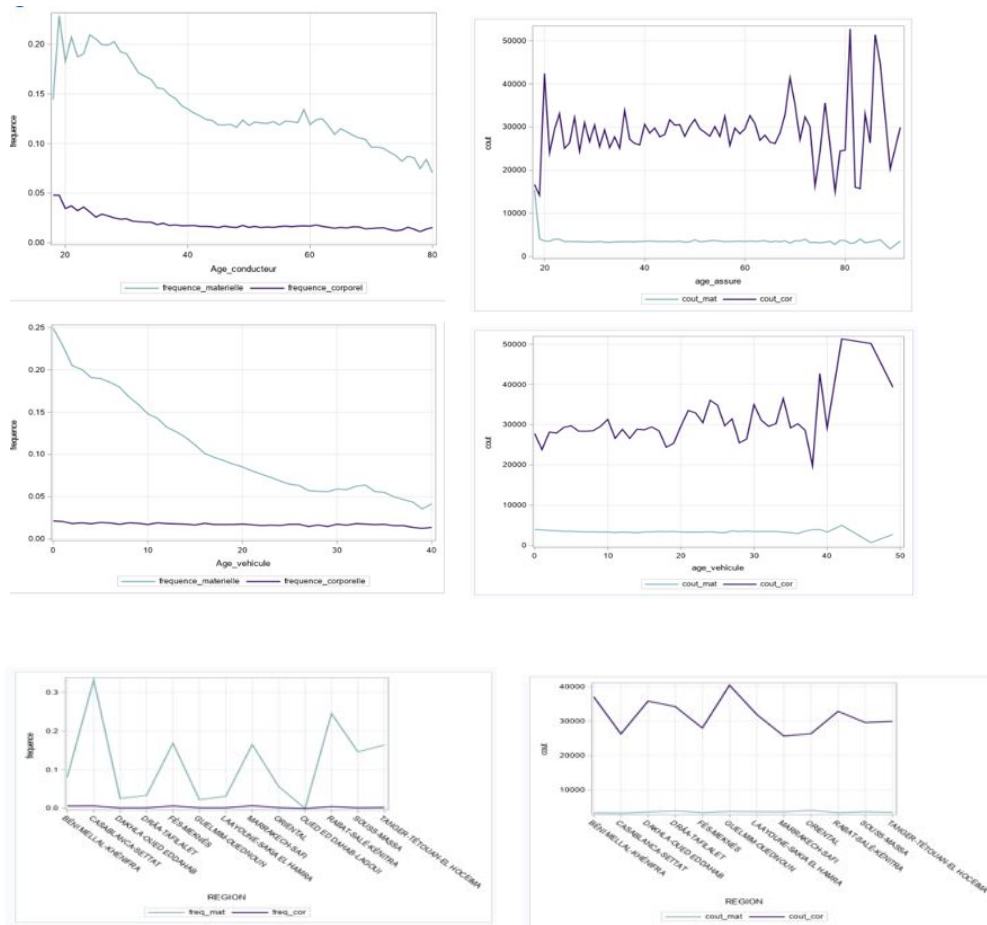


Figure 10: Distribution du cout moyen

Selon la répartition de la charge, il est observé que la plupart des valeurs sont concentrées dans l'intervalle de 0 à 10000dh. L'existence de valeurs très élevées peut être un indicateur de valeurs atypiques et aberrantes.



Les graphiques que nous avons examinés fournissent une justification claire de notre approche séparée des deux segments de la branche Responsabilité Civile (RC), à savoir la RC corporelle et la RC matérielle.

Nous constatons que les résultats obtenus reflètent fidèlement la réalité. Nous pouvons ainsi affirmer que les données sont suffisamment fiables pour poursuivre notre étude.

Par ailleurs, dans l'étape de la tarification, nous traitons séparément la RC corporelle et la RC matérielle.

II.4. Analyse de la corrélation

– Analyse de la corrélation entre les variables quantitatives

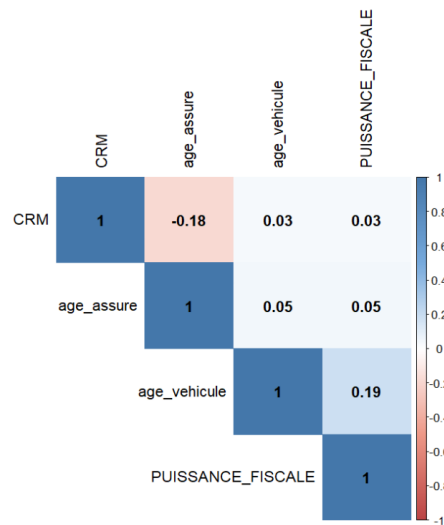


Figure 11: Coefficient de Pearson

La matrice de corrélations confirme l'absence de dépendance entre les quatre variables, ce qui souligne la nécessité de les inclure dans le modèle.

– Analyse de la corrélation entre les variables qualitatives

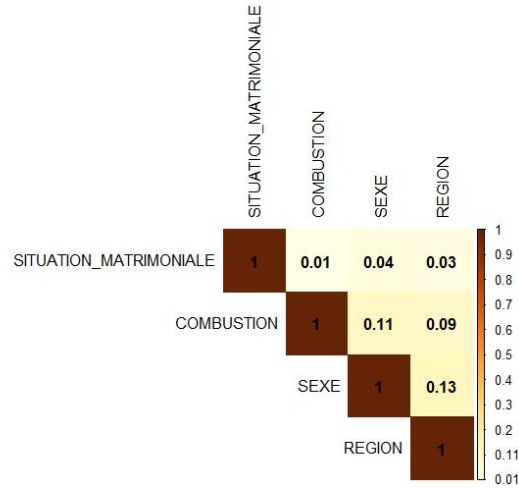


Figure 12: Coefficient V de Cramer

Les résultats obtenus d'indépendance indiquent une indépendance entre chaque paire de variables, étant donné que la valeur du coefficient V de Cramer ne dépasse pas 0.5.

En effet, le coefficient V de Cramer varie de 0 à 1, où 0 indique une absence d'association et 1 indique une association parfaite entre les variables. Plus la valeur de V se rapproche de 1, plus l'association entre les variables est forte.

III. Ecrêtement des sinistres graves

Dans cette partie nous allons pouvoir passer à l'étape de détermination du seuil séparant les sinistres attritionnels et les sinistres graves.

Afin d'accomplir cela, nous utiliserons la théorie des valeurs extrêmes.

– **Loi de Pareto généralisée (GPD):**

La loi de Pareto généralisée $G_{\xi,\sigma}$ est définie par la fonction de répartition :

Pour $\xi \neq 0$:

$$G_{\xi,\sigma}(x) = 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}}$$

Pour $\xi = 0$:

$$G_{\xi,\sigma}(x) = 1 - \exp\left(-\frac{x}{\sigma}\right)$$

La loi de Pareto Généralisée regroupe les lois suivantes selon le paramètre de forme ξ :

- $\xi > 0$, il s'agit de la loi de Pareto simple
- $\xi = 0$, loi exponentielle (limite de $G_{\xi, \sigma}$ lorsque $\xi \rightarrow 0$)
- $\xi < 0$, c'est la loi de Pareto de type II

En effet, si une loi de valeurs extrêmes généralisée (GPD) est bien adaptée pour modéliser les dépassements d'un seuil u_0 , alors les valeurs extrêmes au-delà de ce seuil ($u \geq u_0$) devraient également suivre une loi GPD. Importamment, le paramètre de forme ξ est le même pour les deux distributions, ce qui signifie que le comportement des valeurs extrêmes reste cohérent. Cependant, le paramètre d'échelle σ pour les valeurs extrêmes au-delà du seuil u_0 dépend de la valeur de u . Cette dépendance est donnée par la formule :

$$\sigma(u) = \sigma(u_0) + \xi(u - u_0)$$

σ varie donc linéairement en fonction de u , cette relation n'est valable que si $\xi \neq 0$.

– **Fonction d'excès en moyenne (FME):**

Supposons qu'une loi GPD soit adaptée au problème de dépassement de seuil U_0 d'un échantillon X_1, \dots, X_n , alors on a :

$$E(X - U_0 | X > U_0) = \frac{\sigma_{U_0}}{1 - \xi}, \quad \text{pour } \xi < 1$$

Le modèle doit rester valide pour tout dépassement de seuil $u > U_0$:

$$E(X - u | X > u) = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{U_0}}{1 - \xi} + \xi(u - U_0)$$

Donc la moyenne des excès est une fonction linéaire du seuil u .

Par ailleurs, lorsqu'on analyse les Mean Excess Plot, il faut trouver la valeur du seuil u à partir duquelle la courbe change, devient linéaire et se stabilise.

– **Application:**

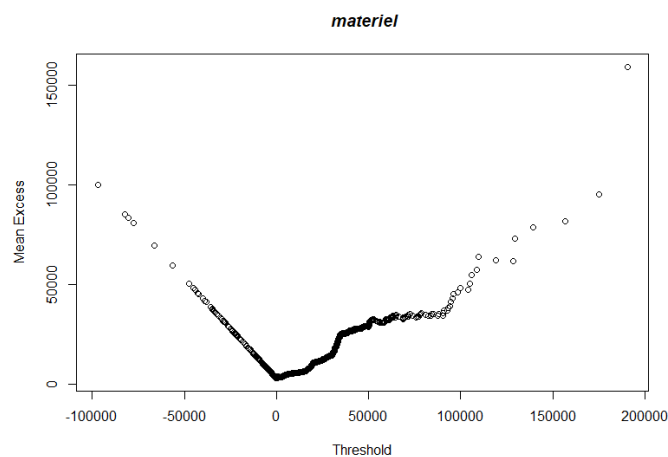


Figure 13: FME pour le segment matériel

Nous remarquons qu'à partir de 50000 la courbe commence à se stabiliser, donc on peut choisir ce seuil pour les charges des sinistres matériels.

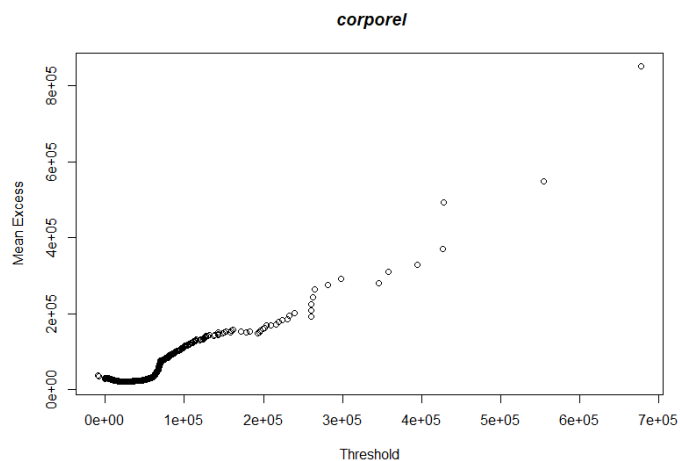


Figure 14: FME pour le segment corporel

D'après le graphe la moyenne de dépassement des seuils, devient affine dans l'intervalle [100000,200000], nous choisissons donc un seuil de 140000.

Après la détermination du seuil des sinistres graves pour les deux segments matériel et corporel, nous pouvons commencer l'étape de la tarification.

En effet, nous allons écrêter les charges à ces seuils, puis ajouter le coût des sinistres graves à la fin, comme montré dans la formule suivante :

$$\begin{aligned} \text{prime pure} &= \text{fréquence matérielle} \times \text{coût matériel} \\ &+ \text{fréquence corporelle} \times \text{coût corporel} \\ &+ \text{coût des larges} \end{aligned}$$

IV. Conclusion

Dans ce chapitre, nous avons présenté les données, effectué des traitements pour garantir leur qualité, et mené des analyses graphiques et de corrélation. Ces étapes ont permis d'identifier les relations clés entre les variables tarifaires, essentielles pour affiner les modèles de tarification. L'écrêtement des sinistres graves a aidé à détecter les charges extrêmes, améliorant ainsi la précision des estimations des coûts moyens.

Cette étude détaillée nous permet d'avoir une meilleure compréhension des éléments qui influencent les risques, ce qui constitue une base solide pour une tarification des primes d'assurance plus précise et juste.

Chapitre 3 : Construction d'un nouveau zonier tarifaire

I. Introduction

La zone géographique est l'un des critères de tarification les plus largement utilisés dans le domaine de l'assurance dommages. En général, les conducteurs qui circulent principalement dans des zones à faible densité urbaine ont moins d'accidents que les autres. Ainsi, la zone géographique dans laquelle un conducteur évolue influence le montant de la prime d'assurance. dans ce chapitre nous visons à améliorer le zonage actuel utilisé dans la tarification automobile en prenant en compte une segmentation plus précise des zones géographiques. Pour ce faire, nous examinons les données sociodémographiques des différentes communes afin de mieux évaluer le risque associé à chacune d'elles. En utilisant ces données externes pertinentes, nous cherchons à définir des groupes de risque homogène qui permettront une tarification plus précise et adaptée aux caractéristiques spécifiques de chaque région.

II.Méthodologie de Travail

L'objectif de la classification des communes dans ce chapitre est double : d'une part, classer les communes en fonction de leurs caractéristiques sociodémographiques similaires, et d'autre part, les classer en fonction de leur comportement en matière de sinistralité. Pour ce faire, une approche supervisée basée sur les arbres de décision a été choisie pour plusieurs raisons :

Les arbres de décision sont une méthode d'apprentissage supervisé qui, contrairement aux méthodes de clustering, prennent en compte une variable à expliquer lors de la modélisation. Ils sont capables de gérer différents types de variables, qu'elles soient quantitatives ou qualitatives, offrant ainsi une grande flexibilité dans l'analyse des données. De plus, les résultats des arbres de décision sont faciles à interpréter car ils organisent les observations sous forme d'arbre, avec des nœuds de décision qui divisent les données en groupes homogènes et des nœuds finaux qui représentent ces groupes. Chaque nœud final est défini par un ensemble de règles claires, ce qui facilite la compréhension et l'analyse des segments créés par l'algorithme.

Pour ce faire , l'algorithme CART a été choisi pour la classification. puisqu'il divise les communes en groupes en se basant à la fois sur leurs caractéristiques explicatives et sur la variable à expliquer, permettant ainsi une classification plus précise et facilement interprétable.

III. Cadre théorique de l'algorithme CART

Les arbres de décision, également connus sous le nom d'Arbres de Régression et de Classification (CART), sont des techniques d'apprentissage statistique largement utilisées pour la régression ou la classification. Ils ont été initialement développés par Leo Breiman et al. en 1984 et sont réputés pour leur efficacité et leur popularité. Les arbres de décision servent de base à de nombreux modèles de prédiction en Data Science, tels que les forêts aléatoires ou le Gradient Boosting Machine. Leur objectif principal est d'expliquer une variable cible à partir de variables explicatives continues ou discrètes, représentées par une matrice X avec m observations et n variables, associée à un vecteur Y à expliquer. Selon la nature de la variable cible Y , les arbres peuvent être utilisés pour la régression (pour les variables numériques) ou pour la classification (pour les variables qualitatives). Les arbres de décision fonctionnent en séparant les observations selon une hiérarchie d'arbre de manière à minimiser une fonction de coût spécifique, telle que la MSE (Mean Squared Error) pour les arbres de régression et le coefficient de GINI pour les arbres de classification.

III.1. Principe de construction de l'arbre

L'arbre de décision commence par un nœud initial, également appelé racine, puis se divise en deux branches menant à des nœuds successifs, jusqu'à ce que l'arbre atteigne une condition d'arrêt. Cette structure crée une série d'emboîtements de rectangles qui délimitent une partition de la population.

Les nœuds terminaux, situés à la base de l'arbre, sont appelés feuilles, ils regroupent des ensembles homogènes d'observations, ces feuilles partagent des combinaisons de modalités de variables explicatives ayant un effet commun sur la variable réponse permettant à la variable d'intérêt de prendre des valeurs aussi homogènes que possible

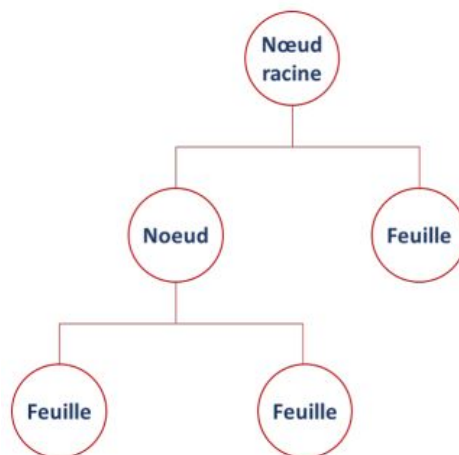


Figure 15: Illustration de l'arbre CART

Notons:

- Y : La variable réponse ;
- p : Le nombre de covariables ;
- X_j : Les covariables, avec $1 \leq j \leq p$;
- π_0 : La quantité que l'on veut prédire.

En général , la quantité que l'on veut prédire est :

$$\pi_0 = E[Y|X = x]$$

On peut également opter pour une autre quantité, telle qu'un quantile , Il faut donc choisir le bon critère de mesure de l'homogénéité des noeuds. Lorsque la quantité visée est l'espérance, la fonction de perte que l'on utilise est l'erreur de généralisation des moindres carrés ou mean squared error (MSE) :

$$E [(\pi(x) - Y)^2]$$

la quantité d'intérêt choisie est donc solution de l'équation suivante :

$$\pi_0(x) = \arg \min_{\pi(x)} E[\varphi(Y, \pi(x))|X = x]$$

avec:

$$\varphi(Y, \pi(x)) = (Y - \pi(x))^2$$

En pratique, les calculs d'espérance sont réalisés de manière empirique. Ainsi, on cherche :

$$\pi_n(x) = \arg \min_{\pi(x)} E[\varphi(Y, \pi(x)) | X = x]$$

posons :

L'espérance empirique :

$$\text{En}(Y) = \frac{1}{n} \sum_{i=1}^n y_i$$

et La variance empirique :

$$V_n(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - \text{En}[Y])^2$$

Pour construire l'arbre, il est nécessaire de diviser chaque nœud en deux nœuds fils, tout en cherchant à minimiser la variance des deux nouveaux nœuds. À chaque nœud construit, le nouvel estimateur de $E[Y]$ devient l'espérance empirique de l'ensemble des observations du nœud en question

Afin de diviser l'ensemble en deux sous-ensembles plus homogènes, nous testons chaque seuil pour chaque variable explicative, et nous sélectionnons le seuil et la variable explicative qui minimisent la variance des deux nouveaux nœuds. Ce processus est répété pour chaque nouveau nœud avec le nouvel ensemble associé, assurant ainsi une segmentation progressive de l'ensemble de données. Ce processus revient à résoudre le problème suivant :

$$\min_{X_i \leq j} (Q_{\text{gauche}} \cdot V(t_{\text{gauche}}) + Q_{\text{droite}} \cdot V(t_{\text{droite}}))$$

avec :

t_{gauche} : le nœud fils de gauche;

t_{droite} : le nœud fils de droite;

Q_{gauche} : la proportion d'individus dans le nœud de gauche;

Q_{droite} : la proportion d'individus dans le nœud de droite.

III.2.l'élagage de l'arbre

L'arbre maximal peut être très grand où chaque variable à expliquer peut résulter dans un nœud séparé. Cependant le découpage doit s'arrêter et donc on doit élaguer l'arbre pour éviter le phénomène de sur-apprentissage ,qui se produit lorsque le modèle d'apprentissage automatique s'ajuste trop précisément aux données d'entraînement, ce qui peut entraîner une performance médiocre

lorsqu'il est confronté à de nouvelles données, Ainsi, l'élagage de l'arbre permet de supprimer les feuilles qui n'apporteraient rien à l'analyse

Soit un arbre T composé de $|T|$ feuilles.
et

$$R(T) = \mathbb{E}_n [\varphi(Y, \pi(x)) | X = x]$$

le taux d'erreur relatif de cet arbre

Sa fonction coût-complexité est égale à :

$$R_\alpha(T) = R(T) + \alpha \cdot |T|$$

Posons :

$$T(\alpha) = \arg \min_{T \leq T_{\max}} R_\alpha(T)$$

$T(\alpha)$ correspond au sous-arbre de l'arbre maximal T_{\max} qui minimise la fonction de complexité en α

Bien évidemment, $T(0) = T_{\max}$

La démarche est relativement simple : nous augmentons itérativement le paramètre α de 0 à $+\infty$, puis pour chaque valeur de α , nous sélectionnons le sous-arbre de T_{\max} qui minimise la fonction de coût-complexité associée à α . Cette procédure génère une série de sous-arbres optimaux de T_{\max} , ainsi qu'une suite de nombres $0 < \alpha_1 < \dots < \alpha_{np}$, où np représente le nombre de sous-arbres créés avant d'atteindre la racine.

III.3. Validation croisée

Dans cette technique, l'ensemble des données est divisé en K ensembles de tailles presque égales. Le premier ensemble est sélectionné comme ensemble de test et le modèle est entraînée sur les autres ensembles $K - 1$

Le taux d'erreur est ensuite calculé après ajustement du modèle aux données de test.

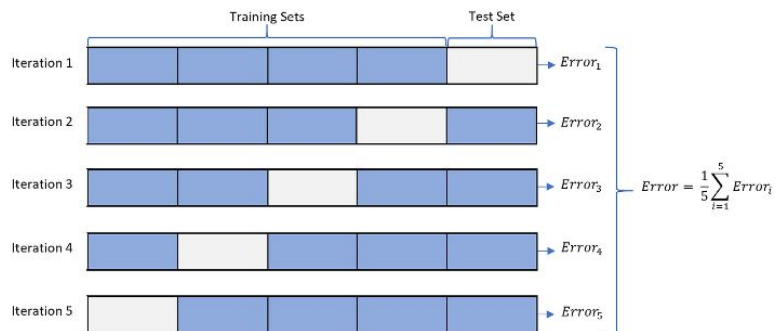


Figure 16: Séparation des données pour la méthode validation croisée du 5-fold

IV. Analyse du zonier actuel

Le zonier actuel, établi par la compagnie en classant les communes en cinq zones de risque en traitant Rabat et Casablanca à part, (Rabat, Casablanca, zone faible, zone moyenne, zone forte), ce découpage repose sur l'expertise des spécialistes quant au comportement présumé de sinistralité dans chaque commune. Cependant, une évaluation de la performance du zonier actuel s'avère nécessaire. Nous examinons donc son évolution sur la période de 2019 à 2022 en utilisant les deux indicateurs de sinistralité retenus, à savoir la fréquence et le coût moyen des sinistres.

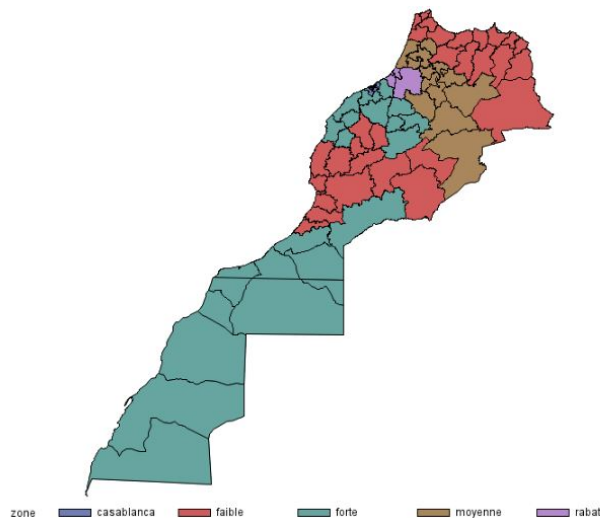


Figure 17: Carte du risque actuel projetée avec SAS

IV.1. Etude des indicateurs de sinistralité

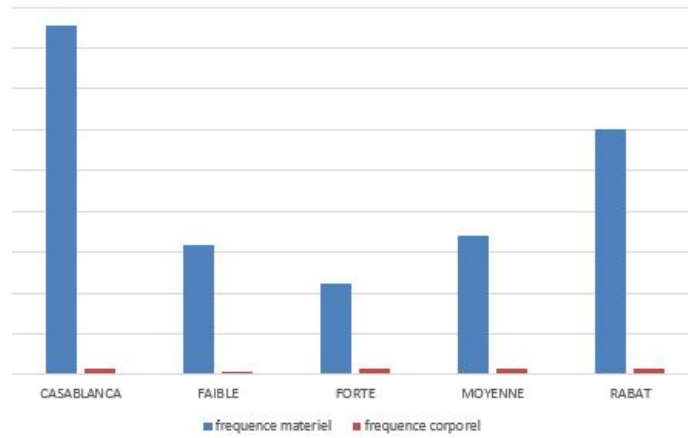


Figure 18: la Fréquence des sinistres

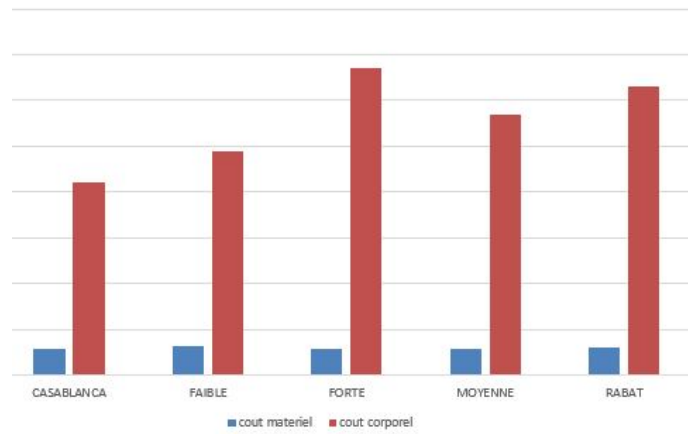


Figure 19: le coût moyen des sinistres

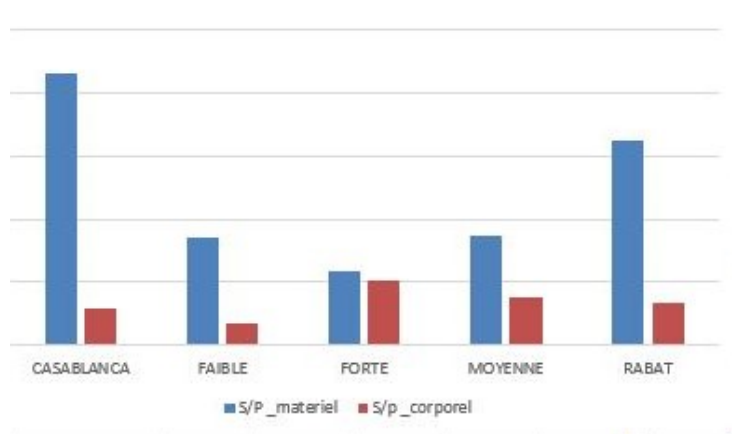


Figure 20: ratio de sinistralité

Le ratio sinistres à primes est un indicateur technique de rentabilité couramment utilisé en assurance non vie, il est égal au rapport de la charge des sinistres divisée par les primes acquises.

À la lumière des trois graphiques, nous observons des disparités dans le découpage actuel, particulièrement en ce qui concerne le segment matériel, en effet au niveau du S/P du segment matériel, nous remarquons que la zone forte, plus risquée que les zones moyenne et faible, est sous tarifée par rapport à ces dernières,

Cela renforce notre démarche visant à explorer de nouvelles méthodes pour élaborer un nouveau zonier, en prenant en compte des critères supplémentaires tels que les caractéristiques sociodémographiques. L'objectif est de parvenir à une segmentation plus précise, par commune, afin d'assurer un maximum de précision.

V.Description de la base de données externe

V.1.description des variables

La base de données externe répertorie les différentes communes de SANLAM avec leurs codes géographiques ainsi que leurs caractéristiques sociodémographiques, fournies sur le site de l'HCP. En effet, nous avons collecté diverses données sociodémographiques considérées comme pertinentes pour expliquer la sinistralité dans les différentes communes du Maroc.

Ces données incluent :

Variable	Signification
Densité	$\frac{\text{nombre total d'habitants}}{\text{superficie de la zone géographique}}$
taux d'activité	$\frac{\text{Population active}}{\text{Population totale en âge de travailler}} \times 100$
prct actives entre 15 et 59	$\frac{\text{Nombre de personnes actives entre 15 et 59 ans}}{\text{Population totale entre 15 et 59 ans}} \times 100$
Taux d'analphabétisme	$\frac{\text{Nombre de personnes analphabètes}}{\text{Population totale}} \times 100$
Taux de chômage	$\frac{\text{Nombre de personnes au chômage}}{\text{Population active}} \times 100$
prct employeurs	$\frac{\text{Nombre d'employeurs}}{\text{Population active}} \times 100$
Prct salariés	$\frac{\text{Nombre de salariés}}{\text{Population active}} \times 100$
prct Indépendants	$\frac{\text{Nombre d'indépendants}}{\text{Population active}} \times 100$
prct travaillant domicile	$\frac{\text{Nombre de personnes travaillant à domicile}}{\text{Population active}} \times 100$
Prct de déplacement en voiture	$\frac{\text{Nombre de personnes se déplaçant en voiture pour aller au travail}}{\text{Population active}} \times 100$

En effet , nous avons fusionné nos données internes portant les caractéristiques des assurés avec la base de données externe sur les caractéristiques des communes . Cette jointure a été réalisée en utilisant la clé "commune", permettant ainsi d'associer les données individuelles des assurés avec les caractéristiques sociodémographiques à leur lieu de résidence.

V.2. Etude des corrélations

À partir de la base de données construite, nous étudions les différentes corrélations entre les variables externes retenues ainsi qu'avec les indicateurs de sinistralité la fréquence des sinistres et le cout moyen , à l'aide du coefficient de Pearson.

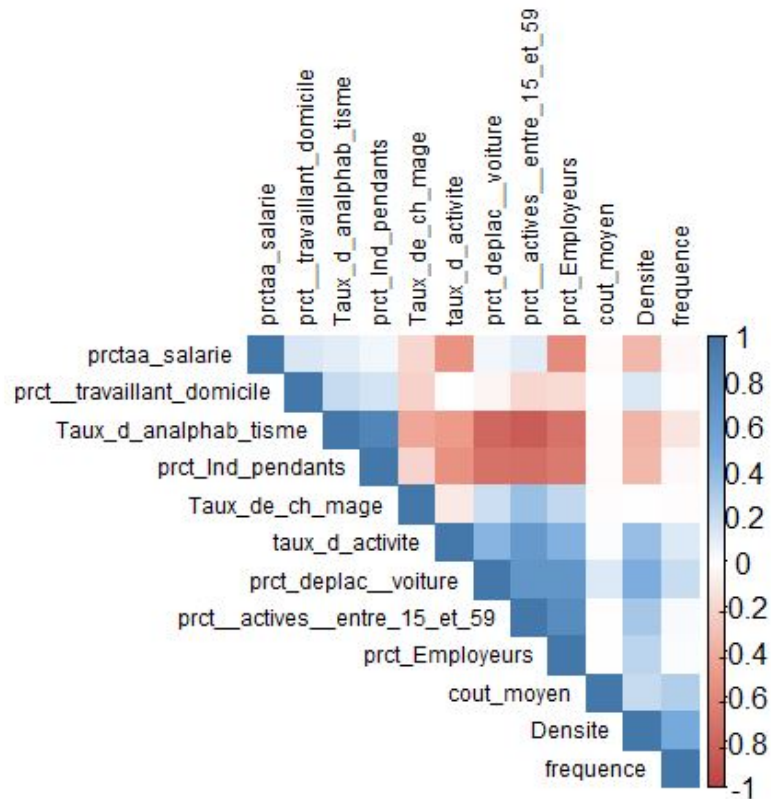


Figure 21: matrice de corrélation

Nous remarquons une corrélation positive entre :

- le taux d'analphabétisme et le pourcentage des indépendants, en effet les personnes analphabètes choisissent fréquemment des emplois indépendants, car les emplois salariés nécessitent des compétences spécifiques qu'ils ne possèdent pas.
- le pourcentage des personnes utilisant une voiture pour se déplacer au travail et le pourcentage des employeurs ,en effet avec des revenus plus

élevés, les employeurs peuvent plus aisément se permettre d'utiliser une voiture pour se rendre au travail.

- le taux d'activité et la densité de la population, en effet les zones peuplées offrent davantage d'opportunités d'emploi, ce qui encourage un plus grand nombre de personnes à s'impliquer dans la vie active.
- la densité de population et la fréquence des sinistres, en raison de la concentration plus élevée de véhicules et de personnes dans les zones densément peuplées.

Nous remarquons une corrélation négative entre :

- le pourcentage des actifs âgés de 15 à 59 ans et le taux d'analphabétisme, en effet L'augmentation du taux d'analphabétisme restreint les possibilités d'emploi, ce qui entraîne une diminution du pourcentage de personnes actives dans ces zones.
- le pourcentage des employeurs et le pourcentage des salariés, en effet les zones avec une forte proportion d'employeurs ont tendance à avoir une proportion plus faible de salariés.
- le taux d'analphabétisme et le pourcentage des personnes utilisant une voiture pour se déplacer au travail, en effet Les analphabètes peuvent avoir moins de possibilités d'accéder aux emplois qui exigent une formation ou des compétences particulières, ce qui les contraint à utiliser des moyens de transport alternatifs, comme les transports en commun ou la marche, pour se rendre au travail.

VI. Application

Par construction l'algorithme CART considère le coût moyen de chaque classe comme la moyenne des coûts moyens, (de même pour la fréquence) ce qui est incorrect et rend cette méthode inadaptée à notre situation; Pour adapter cette approche à notre problème, nous modifions le code de l'algorithme CART en réajustant la valeur moyenne au sein de chaque groupe i comme étant :

$$\bar{y}_{g_i} = \frac{\sum_{j \in g_i} P_j \cdot y_j}{\sum_{j \in g_i} P_j}$$

y_i : la valeur de la variable à expliquer (le coût moyen ou la fréquence)

P_j : le poids de chaque observation j dans le groupe g_i
(l'exposition pour la fréquence et le nombre de sinistres pour le coût moyen)

VI.1. Construction du zonier du cout moyen

VI.1.1. construction de l'arbre maximal

Plusieurs packages existent sous R pour construire des arbres de décision avec l'algorithme CART. Nous avons retenu le package de référence rpart de Therneau et al. (2009) car il nous permet de recoder les modifications à apporter sur l'algorithme. nous construisons à l'aide de la base d'apprentissage un premier arbre de regression (l'arbre maximal ou saturé)

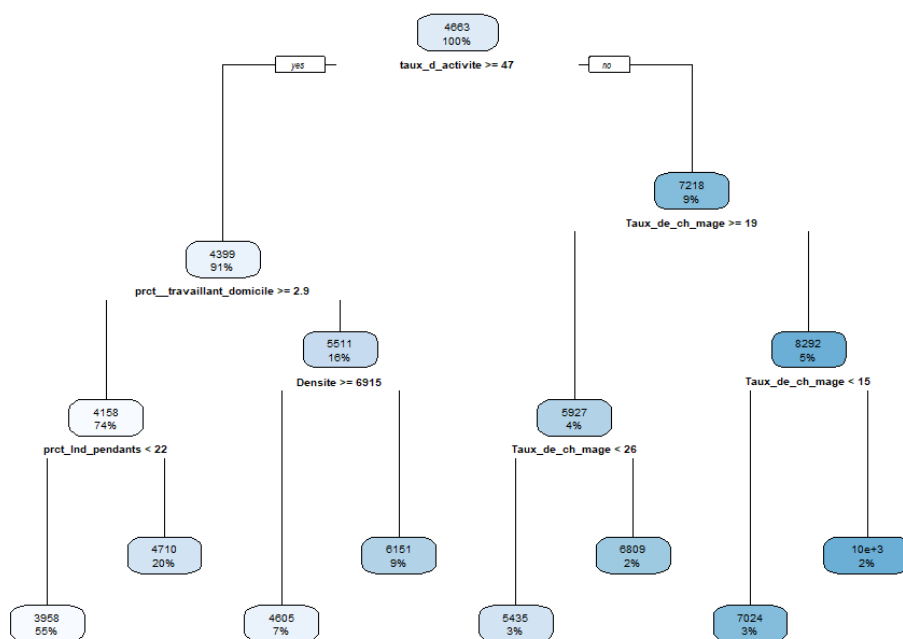


Figure 22: L'arbre de décision maximal pour le coût moyen

VI.1.2 élagage de l'arbre

Pour choisir le bon nombre de feuilles, on procède par validation croisée. La fonction rpart réalise par défaut une estimation des performances de l'arbre par validation croisée à 10 blocs pour chaque niveau de simplification pertinent. On peut afficher les résultats de cette opération grâce à la fonction printcp, comme ci-dessous:

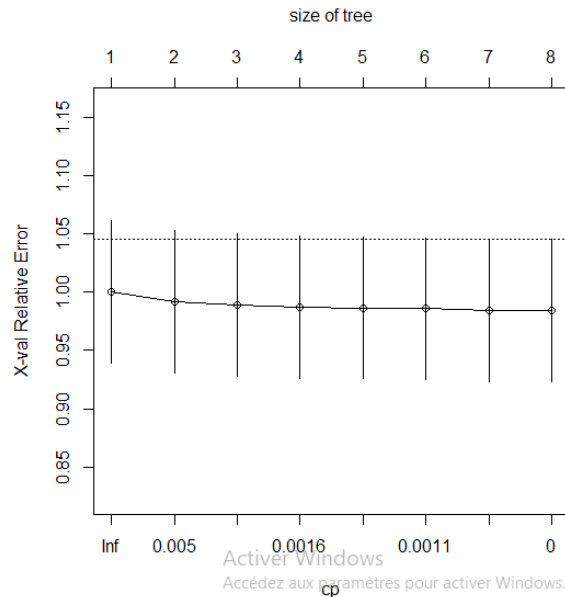


Figure 23: Graphique de l'erreur de la validation croisée

	CP	nsplit	rel error	xerror	xstd
1	0.00841201	0	1.00000	1.00001	0.061242
2	0.00302687	1	0.99159	0.99175	0.061087
3	0.00162025	2	0.98856	0.98912	0.061008
4	0.00161680	3	0.98694	0.98692	0.060990
5	0.00116851	4	0.98532	0.98652	0.060981
6	0.00102175	5	0.98416	0.98574	0.060939
7	0.00023019	6	0.98313	0.98412	0.060818
8	0.00000000	7	0.98290	0.98424	0.060827

Figure 24: la sortie cptable de l'arbre maximal

Pour chaque ligne du tableau correspond à la sortie cptable de l'arbre maximal, nous avons le nombre de divisions nsplit, le cp et les mesures associées (xerror=erreur de validation croisée, xstd=ecart type de l'estimation de l'erreur de validation croisée), en général nous choisissons pour l'élagage de l'arbre optimal le cp tel que l'erreur de validation croisée soit minimale. Nous obtenons donc un arbre avec 7 feuilles.

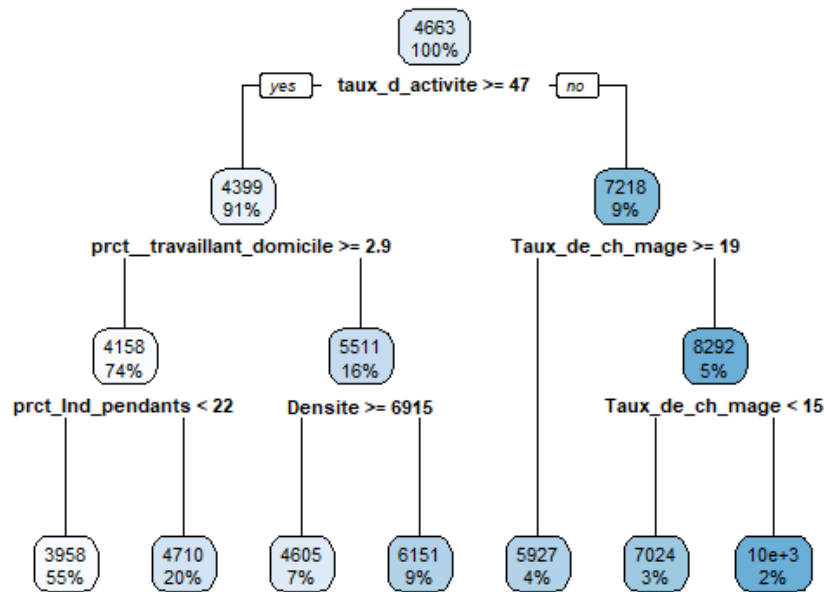


Figure 25: l'arbre optimal pour le coût moyen

L'arbre optimal comporte 7 feuilles, correspondant à 7 classes de zones. Nous constatons que parmi toutes les variables explicatives sociodémographiques sélectionnées, la variable qui permet de réaliser la meilleure séparation binaire de la base de données est le taux d'activité .

Pour identifier les variables les plus influentes dans le modèle, nous allons utiliser la commande `tree$importance` qui permet de quantifier l'importance relative de chaque variable explicative en fonction de sa contribution à la réduction de l'erreur de prédiction

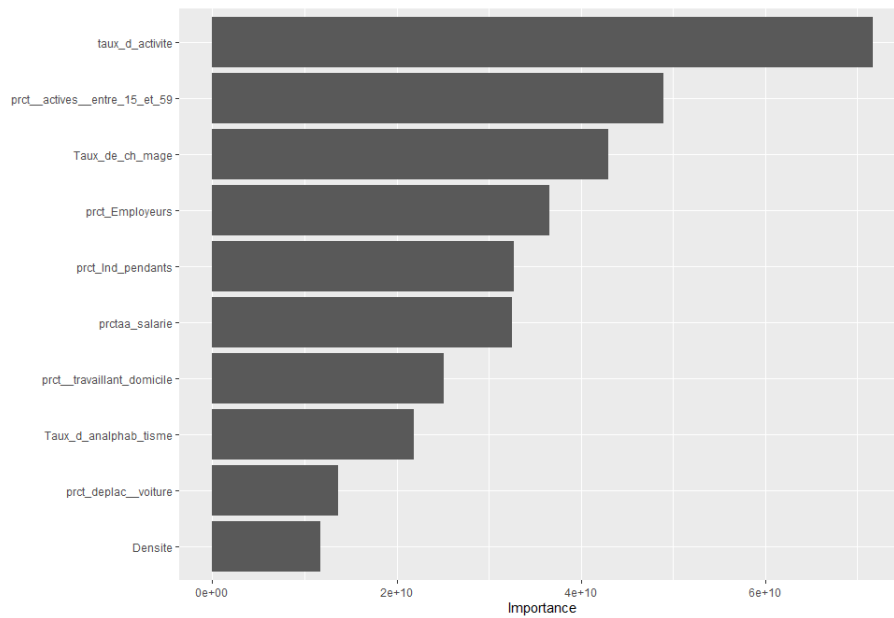


Figure 26: l'importance des variables dans le modèle du cout moyen

Nous constatons que les variables ayant la plus grande importance sont: le taux d'activité et le pourcentage des actifs âgés de 15 à 59 ans. Par conséquent, ces variables sont les plus influentes dans le modèle.

VI.2. Construction du zonier de la fréquence

De la même manière que pour le coût moyen, nous allons construire un arbre maximal en prenant la fréquence comme variable explicative. Ensuite, nous élaguerons l'arbre pour obtenir un arbre optimal. Ainsi, nous obtiendrons un arbre optimal avec 8 feuilles, désignant chacune un segment de zone.

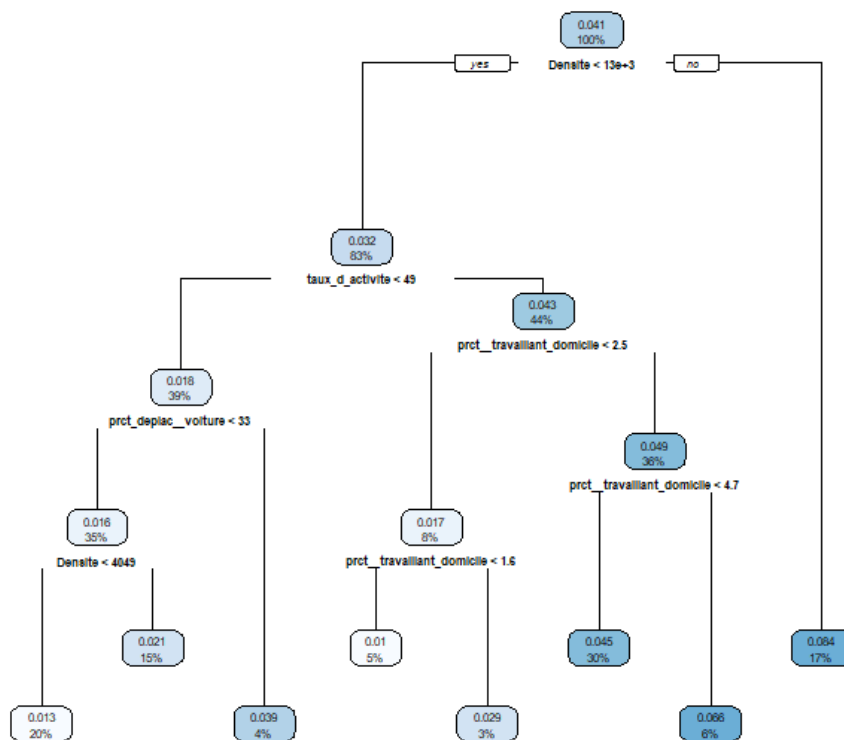


Figure 27: l'arbre optimal pour la fréquence

De même , nous quantifions l'importance des variables dans le modèle CART fréquence .

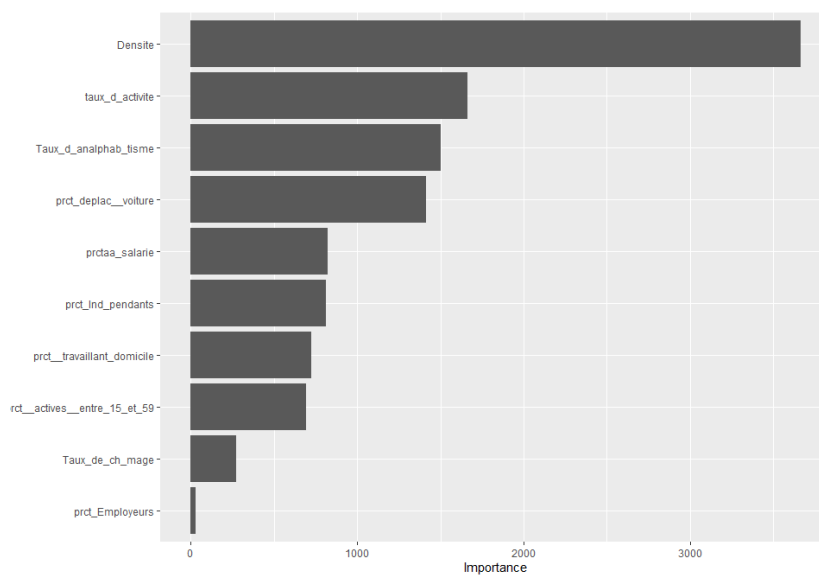


Figure 28: l'importance des variables dans le modèle de la fréquence

D'après le graphe, nous constatons la présence de densité et de taux d'activité au sommet de l'importance des variables, ce qui suggère qu'elles jouent un rôle crucial dans la modélisation et les prédictions effectuées par l'arbre de régression.

VII. Conclusion

Dans ce chapitre, nous avons exploré deux approches de segmentation de zones pour notre analyse. En utilisant la méthode CART, nous avons créé deux ensembles de segments : l'un basé sur la fréquence et l'autre sur le coût moyen. Fascinamment, la segmentation basée sur la fréquence a produit un total de 8 classes de zones distinctes, tandis que celle basée sur le coût moyen en a généré 7. Ces résultats détaillés fournissent une base solide pour notre analyse tarifaire ultérieure. En effet, ces segments nous permettront de mieux comprendre les caractéristiques spécifiques de chaque zone et de concevoir une tarification plus ciblée.

Chapitre 4: Tarification par l'approche classique GLM

I.Introduction

Les modèles linéaires généralisés (GLM), introduits par les statisticiens John Nelder et Robert Wedderburn en 1972, ont profondément influencé la tarification actuarielle. À l'origine, les actuaires utilisaient des modèles linéaires simples pour évaluer les relations entre la valeur des contrats d'assurance et les caractéristiques des risques assurés. Cependant, avec la complexité croissante des problèmes actuariels, ces méthodes sont devenues rapidement insuffisantes. Les GLM ont alors été adoptés pour leur capacité à modéliser des comportements non linéaires et des distributions de résidus non gaussiens. Ces modèles sont devenus incontournables en tarification IARD (Incendie, Accidents et Risques Divers), permettant de capturer les liaisons linéaires entre la variable à expliquer et les variables explicatives, offrant ainsi une meilleure adaptation aux problématiques actuelles en assurance. Dans ce chapitre, nous nous concentrons sur la tarification en utilisant les modèles linéaires généralisés (GLM), en distinguant les segments matériels et corporels. Nous allons adopter deux approches différentes pour le GLM : une en utilisant les segments obtenus par la méthode CART et une autre en utilisant la segmentation actuelle de l'organisme.

NB: En raison de contraintes de taille de mémoire, nous allons présenter les résultats du premier modèle en nous concentrant sur le segment matériel. La même approche sera appliquée à l'autre modèle, et les résultats correspondants seront inclus en annexe.

II.Cadre théorique des modèles linéaires généralisés

Les modèles linéaires généralisés (GLM) se composent de trois éléments :

- La variable de réponse y , qui est une composante aléatoire associée à une distribution de probabilité spécifique.
- La composante déterministe, qui est définie comme une combinaison linéaire des variables explicatives X_1, \dots, X_k .
- La fonction de lien, qui établit la relation fonctionnelle entre la combinaison linéaire des variables X_1, \dots, X_k et l'espérance mathématique de la variable réponse y .

Mathématiquement, cela s'exprime comme suit :

$$g[E(Y)] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

II.1. La composante aléatoire

Soit un n -échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ où les $X_i = (X_{i1}, \dots, X_{ik})$ sont supposées fixes et les Y_i sont des variables aléatoires réelles indépendantes. admettant des distributions issues d'une structure exponentielle. Cela signifie que les lois de ces variables sont dominées par une même mesure dite de référence et que la famille de leurs densités par rapport à cette mesure se met sous la forme :

$$f(Y, \theta, \phi) = \exp \left[\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y, \phi) \right], \quad \forall Y \in S$$

Avec :

- S : un sous-ensemble de \mathbb{N} ou \mathbb{R}
- θ : paramètre canonique
- ϕ : paramètre de dispersion
- a : une fonction définie sur \mathbb{R}^*
- b : une fonction définie sur \mathbb{R} et deux fois dérivable
- c : une fonction définie sur \mathbb{R}^2

Le tableau ci-dessous résume les lois utiles qui sont membres de la famille exponentielle :

II.2 La composante déterministe

La composante déterministe, représentée par une combinaison linéaire

$$\eta(X_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}$$

définit quels sont les prédicteurs.

Certaines variables peuvent être dérivées des variables explicatives incluses dans le modèle. Par exemple :

- $X_j = X_k \cdot X_l$ représente l'interaction entre les variables X_k et X_l
- Ou encore $X_j^2 = X_l$ pour tenir compte de l'effet non linéaire de X_l

L'estimation des paramètres $\beta_0, \beta_1, \dots, \beta_k$ se fait en maximisant la log-vraisemblance du GLM.

<i>Distribution de Y_i</i>	θ	φ	$b(\theta)$	$E(Y_i)$	$Var(Y_i)$
<i>Normale $\sim \mathcal{N}(\mu, \sigma^2)$</i>	μ	σ^2	$\theta^2/2$	μ	σ^2
<i>Poisson $\sim \mathcal{P}(\lambda)$</i>	$\ln(\lambda)$	1	$\exp(\theta)$	λ	λ
<i>Gamma $\sim \Gamma(\mu, \alpha)$</i>	$-1/\mu$	$1/\alpha$	$-\log(-\theta)$	μ	μ/α^2
<i>Binomiale $\sim \mathcal{B}(n, p)/n$</i>	$\log\left(\frac{p}{1-p}\right)$	$\frac{1}{p}$	$\log(1 + \exp(\theta))$	np	$np(1-p)$
<i>Inv. $\sim \mathcal{IG}(\mu, \sigma^2)$</i>	Gauss $-1/2\mu^2$	σ^2	$-(-2\theta)^{1/2}$	μ	μ^3/σ^2

Figure 29: Lois Utiles de la Famille Exponentielle

II.3. La fonction de lien

La fonction de lien traduit la relation fonctionnelle entre la composante aléatoire et la composante déterministe. Elle spécifie le lien entre les deux composantes, plus précisément le lien entre l'espérance conditionnelle de Y_i et la composante déterministe :

$$g(E[Y_i | X_i]) = g(\mu_i) = \eta(X_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}$$

où g est une fonction inversible appelée fonction de lien. Chacune des lois de probabilités de la famille exponentielle possède une fonction de lien spécifique, dite « canonique ».

II.4. Sélection des variables explicatives

✓ **Backward :**

Nous commençons avec le modèle complet, incluant toutes les variables ayant un effet significatif sur le risque à modéliser. Ensuite, nous retirons progressivement la variable la moins significative, celle dont l'élimination entraîne la plus faible augmentation de la déviance.

✓ **forward :**

Nous recherchons la variable la plus significative en termes de déviance. À partir de ce modèle à une seule variable, nous cherchons ensuite la variable qui, associée à la première, explique le mieux le risque, et ainsi de suite

✓ **stepwise :**

Nous ajoutons progressivement des variables au modèle, tout en offrant la possibilité de supprimer une variable précédemment introduite à chaque étape. Cette approche, une combinaison des méthodes backward et forward, est particulièrement utile lorsque les variables explicatives sont corrélées entre elles.

II.5. Significativité des variables

La significativité des coefficients associés aux variables explicatives peut être testée à l'aide du test de Wald. Soit le test suivant :

$$H_0 : \beta_j = 0$$

$$H_1 = \beta_j \neq 0$$

La statistique de Wald s'écrit :

$$W = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

Sous H_0 , la statistique du test suit approximativement une loi Normale $N(0, 1)$.

II.6. Evaluation de la qualité d'ajustement

✓ **La déviance :**

La déviance d'un modèle, notée D , est définie comme une mesure de distance entre l'ajustement de ce modèle et le modèle saturé, le modèle saturé correspond au modèle contenant autant de paramètres que d'observations et fournissant ainsi une description parfaite des données,

$$D = -2(l - l_{\text{sat}})$$

En pratique, on jugera le modèle de mauvaise qualité si:

$$D_{\text{observé}} > \chi_{n-k-1; 1-\alpha}^2$$

où $\chi_{n-k-1; 1-\alpha}^2$ est le quantile d'ordre $1 - \alpha$ de la loi du chi-deux à $n - k - 1$ degrés de liberté, n étant le nombre d'observations et $k + 1$ le nombre de paramètres dans le modèle $(\beta_0, \dots, \beta_k)$.

✓ **Test de Pearson :**

Le χ^2 de Pearson est une autre statistique qui évalue la qualité globale de l'ajustement du modèle. Il est défini comme suit :

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\text{Var}(\hat{\mu}_i)}$$

Elle suit asymptotiquement la loi du χ^2 à $n - k$ degrés de liberté. Ainsi, en réalisant le test asymptotique suivant, nous pouvons évaluer la qualité du modèle :

H_0 : le modèle a k variables est adéquat

H_1 = l'alternative

II.7.Choix du meilleur modèle

✓ **AIC :**

Il est défini par :

$$\text{AIC} = -2l + 2k$$

Où l est la log-vraisemblance maximisée et k le nombre de paramètres dans le modèle. Le meilleur modèle est celui qui possède la plus faible valeur de l'AIC.

✓ **BIC :**

Il est déterminé par :

$$\text{BIC} = -2l + 2k \times \log(n)$$

Où n est le nombre d'observations. Le modèle choisi est celui qui présente la plus petite valeur du BIC.

III.Application

III.1.segmentation des variables

Pour chaque variable concernée, la segmentation a été réalisée en utilisant l'algorithme CART en utilisant la fréquence des sinistres et le coût moyen comme variables dépendantes .

Les tableaux suivants illustrent les résultats obtenus.

La variable	Les classes
Age du conducteur	<31 [31,36[[36,40[[40,70[>=70
Age du véhicule	<8 [8,13[[13,16[[16,21[>=21
CRM	<95 [95,118[[118,125[[125,138[>=138
Puissance fiscale	<7 >=7

Figure 30: La segmentation selon la fréquence des sinistres

La variable	Les classes
Age du conducteur	<24 [24,40[>=40
Age du véhicule	<4 [4,16[[16,20[[20,25[>=25
CRM	<110 [110,125[[125,135[>=135
Puissance fiscale	<7 >=7

Figure 31: La segmentation selon le coût moyen

III.2. Modélisation de la fréquence

III.2.1 Analyse des modèles candidats

Nous disposons de deux lois fondamentales pour modéliser la fréquence des sinistres : la loi de Poisson et la loi binomiale négative. Ci-dessous, nous entreprenons un ajustement graphique de ces lois théoriques (Poisson et binomiale négative) à la loi empirique

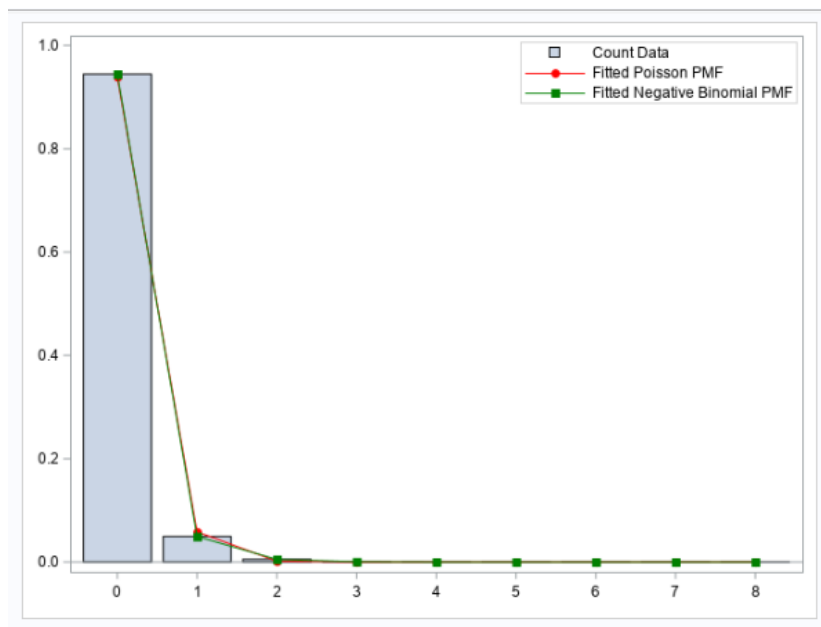


Figure 32: Ajustement de la fréquence des sinistres par la loi de poisson et la loi Binomiale Négative

Nous constatons que la fréquence des sinistres s'ajuste mieux à la loi binomiale négative, pour s'assurer du résultat obtenu nous allons appliquer le GLM avec les deux distributions, en plus, en raison de l'excès de zéros dans la base de données, nous allons retenir en plus de la loi Binomiale Négative, les deux lois ZIP et ZINB pour modéliser la fréquence des sinistres, puis choisir le meilleur modèle, en se basant sur le critère de l'AIC.

III.2.2.mise en oeuvre des modèles

– **Sélection des variables:**

Récapitulatif sur la sélection en avant					
Etape	Effet saisi	DDL	Nombre dans	Khi-2 du score	Pr > khi-2
1	zone_classe_frequenc	7	1	38795.6699	<.0001
2	CRM_freq	4	2	19190.6250	<.0001
3	age_vehicule_freq	4	3	15393.3026	<.0001
4	age_assure_freq	4	4	1843.4649	<.0001
5	SEXE	1	5	609.2543	<.0001
6	combustion_gr	1	6	145.7321	<.0001
7	puissance_freq	1	7	24.1523	<.0001

Figure 33: Les variables sélectionnées par l'approche Forward

D'après le tableau ci-dessus, il est clair que toutes les variables sélectionnées sont significatives selon le test de significativité globale de la variable.

-La loi Binomiale Négative :

Analyse des paramètres estimés du maximum de vraisemblance								
Paramètre		DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%		Khi-2 de Wald	Pr > khi-2
intercept		1	-1.0018	0.0246	-1.0500	-0.9536	1659.46	<.0001
zone_classe_frequenc	1	1	-1.4353	0.0179	-1.4703	-1.4002	6439.65	<.0001
zone_classe_frequenc	2	1	-1.1952	0.0083	-1.2115	-1.1789	20692.2	<.0001
zone_classe_frequenc	3	1	-0.9577	0.0090	-0.9754	-0.9400	11258.7	<.0001
zone_classe_frequenc	4	1	-0.7383	0.0156	-0.7690	-0.7077	2227.63	<.0001
zone_classe_frequenc	5	1	-0.4096	0.0119	-0.4329	-0.3862	1184.83	<.0001
zone_classe_frequenc	6	1	-0.4337	0.0052	-0.4438	-0.4235	6976.52	<.0001
zone_classe_frequenc	7	1	-0.1663	0.0075	-0.1809	-0.1516	495.05	<.0001
zone_classe_frequenc	8	0	0.0000	0.0000	0.0000	0.0000	.	.
CRM_freq	CRM_classe1	1	-1.6462	0.0208	-1.6870	-1.6054	6252.92	<.0001
CRM_freq	CRM_classe2	1	-1.3767	0.0205	-1.4174	-1.3360	4391.39	<.0001
CRM_freq	CRM_classe3	1	-0.3991	0.0216	-0.4413	-0.3568	342.74	<.0001
CRM_freq	CRM_classe4	1	-0.7583	0.0352	-0.8273	-0.6894	464.51	<.0001
CRM_freq	CRM_classe5	0	0.0000	0.0000	0.0000	0.0000	.	.
age_vehicule_freq	age_vehicule_classe1	1	0.8849	0.0077	0.8699	0.9000	13282.6	<.0001
age_vehicule_freq	age_vehicule_classe2	1	0.6731	0.0080	0.6575	0.6887	7162.04	<.0001
age_vehicule_freq	age_vehicule_classe3	1	0.5266	0.0103	0.5065	0.5467	2631.38	<.0001
age_vehicule_freq	age_vehicule_classe4	1	0.3290	0.0093	0.3109	0.3472	1264.12	<.0001
age_vehicule_freq	age_vehicule_classe5	0	0.0000	0.0000	0.0000	0.0000	.	.
age_assure_freq	age_assure_classe1	1	0.7144	0.0131	0.6887	0.7401	2969.29	<.0001
age_assure_freq	age_assure_classe2	1	0.5482	0.0129	0.5230	0.5734	1814.80	<.0001
age_assure_freq	age_assure_classe3	1	0.4403	0.0131	0.4146	0.4659	1133.22	<.0001
age_assure_freq	age_assure_classe4	1	0.2658	0.0119	0.2424	0.2891	497.53	<.0001
age_assure_freq	age_assure_classe5	0	0.0000	0.0000	0.0000	0.0000	.	.
SEXE	F	1	0.0844	0.0058	0.0730	0.0958	209.99	<.0001
SEXE	M	0	0.0000	0.0000	0.0000	0.0000	.	.
combustion_gr	classe1	1	-0.1334	0.0060	-0.1451	-0.1217	499.83	<.0001
combustion_gr	classe2	0	0.0000	0.0000	0.0000	0.0000	.	.
puissance_freq	puissance_classe1	1	-0.0307	0.0046	-0.0368	-0.0247	44.58	<.0001
puissance_freq	puissance_classe2	0	0.0000	0.0000	0.0000	0.0000	.	.

Figure 34: Estimation des paramètres du modèle Binomial Négatif

Obs.	Criterion	DF	Value	ValueDF	pvalue
1	Ecart	36E5	1091141.1264	0.3033	1
2	Déviante normalisée	36E5	1091141.1264	0.3033	1
3	AIC (préférer les petites valeurs)	—	1504823.7532	—	.
4	BIC (préférer les petites valeurs)	—	1505138.0505	—	.

Figure 35: Evaluation de la qualité d'ajustement du GLM pour la loi Binomiale Négative

—La loi de poisson :

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre		DDL	Estimation	Erreur type	Intervalle de confiance de Valid à 95%	Khi-2 de Valid	Pr > khi-2
Intercept		1	-1.0018	0.0246	-1.0500 -0.9536	166.08	< 0.001
zone_classe_frequenc	1	1	-1.4353	0.0179	-1.4703 -1.4002	6439.87	< 0.001
zone_classe_frequenc	2	1	-1.1952	0.0083	-1.2115 -1.1789	20697.0	< 0.001
zone_classe_frequenc	3	1	-0.9577	0.0090	-0.9754 -0.9400	11261.6	< 0.001
zone_classe_frequenc	4	1	-0.7383	0.0156	-0.7690 -0.7077	2227.79	< 0.001
zone_classe_frequenc	5	1	-0.4096	0.0119	-0.4329 -0.3862	1184.93	< 0.001
zone_classe_frequenc	6	1	-0.4337	0.0052	-0.4438 -0.4235	6980.62	< 0.001
zone_classe_frequenc	7	1	-0.1663	0.0075	-0.1809 -0.1516	495.07	< 0.001
zone_classe_frequenc	8	0	0.0000	0.0000	0.0000	.	.
CRM_freq	CRM_classe1	1	-1.6462	0.0208	-1.6870 -1.6054	6255.75	< 0.001
CRM_freq	CRM_classe2	1	-1.3767	0.0208	-1.4174 -1.3360	4393.84	< 0.001
CRM_freq	CRM_classe3	1	-0.3991	0.0216	-0.4413 -0.3568	342.77	< 0.001
CRM_freq	CRM_classe4	1	-0.7583	0.0352	-0.8273 -0.6894	464.55	< 0.001
CRM_freq	CRM_classe5	0	0.0000	0.0000	0.0000	.	.
age_vehicule_freq	age_vehicule_classe1	1	0.8849	0.0077	0.8699 0.9000	13284.2	< 0.001
age_vehicule_freq	age_vehicule_classe2	1	0.6731	0.0080	0.6575 0.6887	7162.07	< 0.001
age_vehicule_freq	age_vehicule_classe3	1	0.6266	0.0103	0.5065 0.5467	2631.40	< 0.001
age_vehicule_freq	age_vehicule_classe4	1	0.3290	0.0093	0.3109 0.3472	1264.14	< 0.001
age_vehicule_freq	age_vehicule_classe5	0	0.0000	0.0000	0.0000	.	.
age_assure_freq	age_assure_classe1	1	0.7144	0.0131	0.6887 0.7401	2969.75	< 0.001
age_assure_freq	age_assure_classe2	1	0.5482	0.0129	0.5230 0.5734	1814.91	< 0.001
age_assure_freq	age_assure_classe3	1	0.4403	0.0131	0.4146 0.4659	1133.26	< 0.001
age_assure_freq	age_assure_classe4	1	0.2658	0.0119	0.2424 0.2891	497.54	< 0.001
age_assure_freq	age_assure_classe5	0	0.0000	0.0000	0.0000	.	.
SEXE	F	1	0.0844	0.0058	0.0730 0.0958	209.95	< 0.001
SEXE	M	0	0.0000	0.0000	0.0000	.	.
combustion_gr	classe1	1	-0.1334	0.0060	-0.1451 -0.1217	499.84	< 0.001
combustion_gr	classe2	0	0.0000	0.0000	0.0000	.	.
puissance_freq	puissance_classe1	1	-0.0307	0.0046	-0.0398 -0.0217	44.58	< 0.001
puissance_freq	puissance_classe2	0	0.0000	0.0000	0.0000	.	.

Figure 36: Estimation des paramètres du modèle de poisson

Obs.	Criterion	DF	Value	ValueDF	pvalue
1	Ecart	30E5	1091156.2847	0.3033	1
2	Déviante normalisée	30E5	1091156.2847	0.3033	1
3	AIC (préférer les petites valeurs)	-	1504823.2340	-	.
4	BIC (préférer les petites valeurs)	-	1505124.4356	-	.

Figure 37: Evaluation de la qualité d'ajustement du GLM pour la loi de poisson

-La loi ZIP :

Paramètre		DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%	Khi-2 de Wald	Pr > khi-2
Intercept		1	-0.5786	0.0269	-0.6312 -0.5259	463.63	<.0001
zone_classe_frequenc	1	1	-1.4410	0.0161	-1.4764 -1.4055	6347.61	<.0001
zone_classe_frequenc	2	1	-1.2020	0.0085	-1.2196 -1.1854	20135.2	<.0001
zone_classe_frequenc	3	1	-0.9660	0.0092	-0.9841 -0.9479	10952.1	<.0001
zone_classe_frequenc	4	1	-0.7451	0.0160	-0.7765 -0.7138	2166.32	<.0001
zone_classe_frequenc	5	1	-0.4134	0.0123	-0.4374 -0.3893	1131.50	<.0001
zone_classe_frequenc	6	1	-0.4401	0.0054	-0.4507 -0.4294	6598.70	<.0001
zone_classe_frequenc	7	1	-0.1651	0.0079	-0.1805 -0.1497	441.87	<.0001
zone_classe_frequenc	8	0	0.0000	0.0000	0.0000 0.0000	.	.
CRM_freq	CRM_classe1	1	-1.6268	0.0230	-1.6718 -1.5817	5010.21	<.0001
CRM_freq	CRM_classe2	1	-1.3614	0.0229	-1.4063 -1.3164	3522.74	<.0001
CRM_freq	CRM_classe3	1	-0.4145	0.0238	-0.4610 -0.3679	304.25	<.0001
CRM_freq	CRM_classe4	1	-0.7401	0.0379	-0.8144 -0.6657	380.31	<.0001
CRM_freq	CRM_classe5	0	0.0000	0.0000	0.0000 0.0000	.	.
age_vehicule_freq	age_vehicule_classe1	1	0.8892	0.0076	0.8738 0.9046	12669.3	<.0001
age_vehicule_freq	age_vehicule_classe2	1	0.6736	0.0081	0.6577 0.6895	6889.91	<.0001
age_vehicule_freq	age_vehicule_classe3	1	0.5268	0.0105	0.5062 0.5473	2517.44	<.0001
age_vehicule_freq	age_vehicule_classe4	1	0.3286	0.0094	0.3101 0.3470	1218.63	<.0001
age_vehicule_freq	age_vehicule_classe5	0	0.0000	0.0000	0.0000 0.0000	.	.
age_assure_freq	age_assure_classe1	1	0.7243	0.0135	0.6978 0.7508	2863.46	<.0001
age_assure_freq	age_assure_classe2	1	0.5518	0.0133	0.5258 0.5778	1728.88	<.0001
age_assure_freq	age_assure_classe3	1	0.4440	0.0135	0.4175 0.4704	1083.76	<.0001
age_assure_freq	age_assure_classe4	1	0.2682	0.0123	0.2442 0.2922	478.41	<.0001
age_assure_freq	age_assure_classe5	0	0.0000	0.0000	0.0000 0.0000	.	.
SEXE	F	1	0.0909	0.0061	0.0789 0.1029	220.16	<.0001
SEXE	M	0	0.0000	0.0000	0.0000 0.0000	.	.
combustion_gr	classe1	1	-0.1330	0.0062	-0.1451 -0.1209	464.96	<.0001
combustion_gr	classe2	0	0.0000	0.0000	0.0000 0.0000	.	.
puissance_freq	puissance_classe1	1	-0.0343	0.0048	-0.0437 -0.0249	51.30	<.0001
puissance_freq	puissance_classe2	0	0.0000	0.0000	0.0000 0.0000	.	.
Echelle		0	1.0000	0.0000	1.0000 1.0000	.	.

Figure 38: Estimation des paramètres de la loi ZIP

Obs	Criterion	DF	Value	ValueDF
1	Ecart	_	1499568.6431	0.4165
2	Déviante normalisée	_	1499568.6431	0.4165
3	AIC (préférer les petites valeurs)	_	1499617.6431	_
4	BIC (préférer les petites valeurs)	_	1499931.9404	_

Figure 39: Evaluation de la qualité d'ajustement du GLM pour la loi ZIP

-La loi ZINB :

Analyse de paramètres estimés du maximum de vraisemblance							
Paramètre		DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%	Khi-2 de Wald	Pr > khi-2
Intercept		1	-0.9605	0.0286	-1.0166 -0.9045	1129.38	<.0001
zone_classe_frequenc	1	1	-1.4470	0.0182	-1.4827 -1.4114	6335.59	<.0001
zone_classe_frequenc	2	1	-1.2095	0.0086	-1.2263 -1.1927	20010.2	<.0001
zone_classe_frequenc	3	1	-0.9739	0.0093	-0.9922 -0.9556	10881.4	<.0001
zone_classe_frequenc	4	1	-0.7529	0.0162	-0.7846 -0.7212	2162.69	<.0001
zone_classe_frequenc	5	1	-0.4202	0.0125	-0.4447 -0.3958	1134.31	<.0001
zone_classe_frequenc	6	1	-0.4469	0.0055	-0.4578 -0.4361	6527.20	<.0001
zone_classe_frequenc	7	1	-0.1694	0.0080	-0.1852 -0.1537	445.27	<.0001
zone_classe_frequenc	8	0	0.0000	0.0000	0.0000 0.0000	.	.
CRM_freq	CRM_classe1	1	-1.6842	0.0252	-1.7335 -1.6348	4481.78	<.0001
CRM_freq	CRM_classe2	1	-1.4207	0.0251	-1.4700 -1.3715	3198.73	<.0001
CRM_freq	CRM_classe3	1	-0.4098	0.0260	-0.4608 -0.3587	247.48	<.0001
CRM_freq	CRM_classe4	1	-0.7808	0.0399	-0.8591 -0.7026	382.18	<.0001
CRM_freq	CRM_classe5	0	0.0000	0.0000	0.0000 0.0000	.	.
age_vehicule_freq	age_vehicule_classe1	1	0.8947	0.0079	0.8792 0.9102	12777.5	<.0001
age_vehicule_freq	age_vehicule_classe2	1	0.6738	0.0082	0.6577 0.6899	6763.53	<.0001
age_vehicule_freq	age_vehicule_classe3	1	0.5258	0.0106	0.5050 0.5466	2455.28	<.0001
age_vehicule_freq	age_vehicule_classe4	1	0.3271	0.0095	0.3095 0.3457	1187.97	<.0001
age_vehicule_freq	age_vehicule_classe5	0	0.0000	0.0000	0.0000 0.0000	.	.
age_assure_freq	age_assure_classe1	1	0.7304	0.0135	0.7034 0.7573	2820.95	<.0001
age_assure_freq	age_assure_classe2	1	0.5575	0.0135	0.5311 0.5839	1710.17	<.0001
age_assure_freq	age_assure_classe3	1	0.4474	0.0137	0.4205 0.4742	1067.08	<.0001
age_assure_freq	age_assure_classe4	1	0.2700	0.0124	0.2456 0.2944	470.64	<.0001
age_assure_freq	age_assure_classe5	0	0.0000	0.0000	0.0000 0.0000	.	.
SEXE	F	1	0.0956	0.0063	0.0833 0.1078	232.20	<.0001
SEXE	M	0	0.0000	0.0000	0.0000 0.0000	.	.
combustion_gr	classe1	1	-0.1354	0.0063	-0.1476 -0.1231	467.60	<.0001
combustion_gr	classe2	0	0.0000	0.0000	0.0000 0.0000	.	.
pulsance_freq	pulsance_classe1	1	-0.0342	0.0049	-0.0437 -0.0246	49.16	<.0001
pulsance_freq	pulsance_classe2	0	0.0000	0.0000	0.0000 0.0000	.	.

Figure 40: Estimation des paramètres de la loi ZINB

Obs.	Criterion	DF	Value	ValueDF
1	Ecart	_	1497733.4433	0.4160
2	Déviante normalisée	_	1497733.4433	0.4160
3	AIC (préférer les petites valeurs)	_	1497783.4433	_
4	BIC (préférer les petites valeurs)	_	1498110.8363	_

Figure 41: Evaluation de la qualité d'ajustement du GLM pour la loi ZINB

D'après les résultats obtenus, nous remarquons que pour les quatre modèles, toutes nos variables tarifaires sont significatives (test de Wald).

Ainsi, la valeur de la p-value du test d'ajustement basé sur la déviance est supérieure à 5%, donc on accepte H_0 en disant que les modèles s'ajustent convenablement aux données.

Pour choisir le meilleur modèle, nous allons comparer les AIC des différents modèles et retenir celui ayant la valeur minimale. le modèle le plus pertinent est celui de **ZINB**.

III.2.3. Analyse graphique des résidus

Les résidus permettent d'évaluer la pertinence du modèle en comparant les valeurs de la variable à expliquer et ses estimations. Nous allons examiner deux catégories de résidus : Les résidus bruts (Raw residuals) et Les résidus de Pearson.

Si les résidus observés sont autour de l'axe des abscisses et avec une variance constante, le modèle est considéré comme valide.

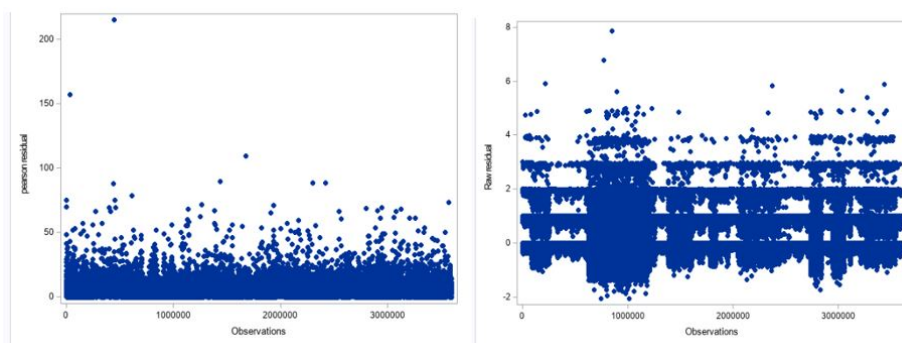


Figure 42: analyse des résidus pour le modèle ZINB

Nous remarquons que le modèle ZINB est valide pour la fréquence des sinistres. Effectivement, la majorité des résidus de Pearson se trouvent

autour de l'axe des abscisses, ce qui indique que le modèle offre une bonne prédiction.

En revanche, quelques points éloignés de la bande horizontale pourraient être attribuables à des valeurs aberrantes.

III.3. Modélisation du coût moyen

III.3.1. Analyse des modèles candidats

Avant de passer au GLM, il est crucial de faire le bon choix de la loi qui ajuste mieux le coût moyen. Afin d'accomplir cela, nous analysons à l'aide des QQplot les distributions possibles qui peuvent l'ajuster: gamma et log-normal

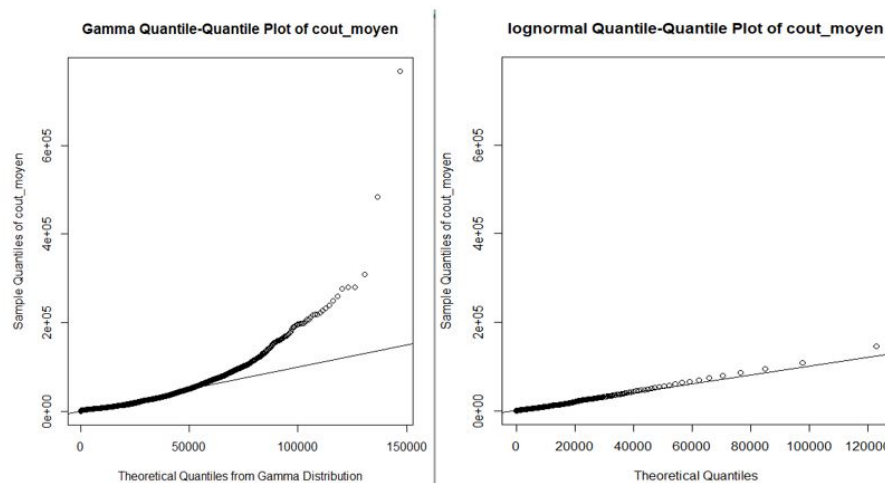


Figure 43: QQplot des lois Gamma et Log-Normale

Graphiquement, nous constatons que la loi log-normale s'ajuste mieux aux données, car l'alignement avec la bissectrice est meilleur.

Pour confirmer ce constat nous allons comparer la qualité d'ajustement du GLM pour la loi gamma et log-normale

III.3.2. mise en oeuvre des modèles

– Sélection des variables:

Les variables sélectionnées par l'approche Forward sont représentées dans le tableau suivant :

Note: All effects have been entered into the model.

Récapitulatif sur la sélection en avant					
Etape	Effet saisi	DDL	Nombre dans	Khi-2 du score	Pr > khi-2
1	age_assure_cout	2	1	6324.1555	<.0001
2	age_vehicule_cout	4	2	4372.8116	<.0001
3	SEXE	1	3	488.3377	<.0001
4	combustion_gr	1	4	346.7392	<.0001
5	CRM_cout_gr	1	5	51.2580	<.0001
6	zone_classe_cout_gn	2	6	55.0072	<.0001
7	puissance_cout	1	7	48.2032	<.0001

Figure 44: sélection des variables par l'approche Forward

- D'après la sortie de la méthode forward, nous constatons que toutes les variables sont sélectionnées dans le modèle du coût moyen.

– **Choix du meilleur modèle :**

- La première implémentation des deux modèles montre que les modalités de certaines variables ne sont pas significatives (les p-values de la colonne du test de Wald sont supérieures à 5%), ce qui a nécessité le regroupement de certaines modalités.

La qualité d'ajustement du GLM pour les lois gamma et log-normale après regroupement est présentée dans les tableaux ci-dessous.

Obs.	Criterion	DF	Value	ValueDF	pvalue
1	Ecart	17E4	230510.4597	1.3908	0
2	Déviance normalisée	17E4	195290.5279	1.1783	0
7	AIC (préférer les petites valeurs)	-	3104222.6766	-	.
9	BIC (préférer les petites valeurs)	-	3104372.9508	-	.

Gamma

Obs.	Criterion	DF	Value	ValueDF	pvalue
1	Ecart	17E4	232767.4667	1.4044	0.00000
2	Déviance normalisée	17E4	165757.0000	1.0001	0.49122
7	AIC (préférer les petites valeurs)	-	526701.3245	-	.
9	BIC (préférer les petites valeurs)	-	526831.5621	-	.

log-normal

Figure 45: Evaluation de la qualité d'ajustement du GLM pour la loi gamma et log-normale

- D'après les résultats obtenus, il semble bien que la loi **log-normale** soit le modèle jugé le plus adéquat, ayant ainsi l'AIC minimal.

- L'estimation des paramètres pour la loi log-normale est présentée dans le tableau ci-dessous:

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre		DDL	Estimation	Erreur type	Intervalle de confiance de Wald à 95%	Khi-2 de Wald	Pr > khi-2
Intercept		1	7.8491	0.0137	7.8222 7.8760	329811	<.0001
combustion_gr	classe1	1	-0.0329	0.0081	-0.0489 -0.0169	16.32	<.0001
combustion_gr	classe2	0	0.0000	0.0000	0.0000 0.0000	.	.
age_assure_cout_gr	age_assure_classe1	1	0.1534	0.0251	0.1041 0.2027	37.23	<.0001
age_assure_cout_gr	age_assure_classe2	0	0.0000	0.0000	0.0000 0.0000	.	.
age_vehicule_cout	age_vehicule_classe1	1	-0.1698	0.0103	-0.1870 -0.1467	263.43	<.0001
age_vehicule_cout	age_vehicule_classe2	1	-0.1590	0.0107	-0.1800 -0.1381	221.37	<.0001
age_vehicule_cout	age_vehicule_classe3	1	-0.1219	0.0140	-0.1493 -0.0945	75.85	<.0001
age_vehicule_cout	age_vehicule_classe4	1	-0.0838	0.0125	-0.1082 -0.0593	44.96	<.0001
age_vehicule_cout	age_vehicule_classe5	0	0.0000	0.0000	0.0000 0.0000	.	.
CRM_cout_gr	CRM_classe1&3	1	0.0272	0.0097	0.0083 0.0462	7.92	0.0049
CRM_cout_gr	CRM_classe2&4	0	0.0000	0.0000	0.0000 0.0000	.	.
zone_classe_cout_gn	1	1	0.1606	0.0245	0.1126 0.2085	43.01	<.0001
zone_classe_cout_gn	6,2,3	1	-0.0367	0.0070	-0.0503 -0.0231	27.85	<.0001
zone_classe_cout_gn	7,4,5	0	0.0000	0.0000	0.0000 0.0000	.	.
puissance_cout	puissance_classe1	1	-0.0483	0.0084	-0.0808 -0.0359	57.86	<.0001
puissance_cout	puissance_classe2	0	0.0000	0.0000	0.0000 0.0000	.	.
SEXE	F	1	-0.0373	0.0080	-0.0529 -0.0216	21.74	<.0001
SEXE	M	0	0.0000	0.0000	0.0000 0.0000	.	.
Echelle		1	1.1850	0.0021	1.1810 1.1891		

Figure 46: L'estimation des paramètres pour la loi log-normale

III.3.3. Analyse graphique des résidus

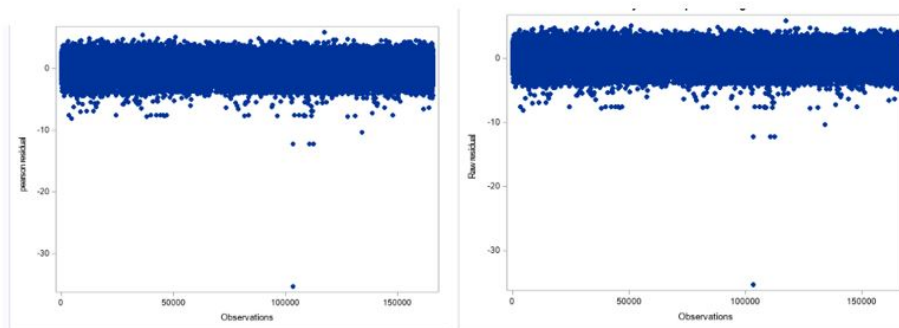


Figure 47: Analyse graphique des résidus pour le modèle Log-Normal

L'examen des résidus révèle une distribution aléatoire autour de zéro, ce qui indique que le modèle linéaire est validé.

De plus, les résidus forment une bande horizontale autour de zéro, suggérant une variance constante des erreurs.

IV. Conclusion

Dans ce chapitre, nous avons abordé la partie théorique des modèles linéaires généralisés (GLM), en exposant en détail leurs bases mathématiques et leurs utilisations potentielles. Ensuite, nous avons mis en pratique cette théorie pour la responsabilité civile matérielle en utilisant une segmentation réalisée à l'aide de l'algorithme CART afin de représenter la fréquence des sinistres et le coût moyen.

Selon nos analyses, le modèle ZINB a démontré sa validité pour la fréquence des sinistres, tandis que la loi log-normale est approuvée pour le coût moyen. Finalement, nous avons l'intention de reproduire ce même travail en utilisant la segmentation suggérée par l'organisme, dans le but de comparer les performances et la pertinence des différentes approches dans la modélisation des données de sinistres.

Chapitre 5: Tarification par des méthodes d'apprentissage

I.Introduction

Les modèles de GLM sont largement utilisés dans la tarification, tant en assurance automobile qu'en assurance santé. Ces modèles se distinguent par leur capacité à détecter les effets non linéaires et à prendre en compte la nature non gaussienne des distributions de résidus. Cependant, malgré leurs performances supérieures par rapport aux modèles de régression classiques, les contraintes qu'ils imposent, telles que les interactions entre variables explicatives et la structure du risque, peuvent parfois conduire à des résultats non pertinents.

Dans ce chapitre, nous proposons d'appliquer les méthodes d'apprentissage statistique XGBoost et CART afin de prédire séparément la fréquence et le cout moyen Nous comparerons ensuite les résultats obtenus avec ceux de notre modèle de référence, le GLM.

NB : De même, par contrainte de taille, nous allons présenter uniquement les résultats de la responsabilité civile matérielle. La même approche sera appliquée pour le segment corporel.

II.Cadre théorique de l'algorithme XGBOOST

Soit $\{(x_i, y_i)\}_{i=1}^n$ un ensemble de données où x_i est le vecteur des caractéristiques pour l'observation i et y_i est la variable cible. L'objectif est de prédire y_i en minimisant une fonction de perte L . La fonction de perte pour une observation est souvent une fonction de l'erreur entre la prédiction et la valeur réelle :

$$L(y_i, \hat{y}_i)$$

avec :

- \hat{y}_i : la prédiction pour l'observation i

À chaque étape t , un nouvel arbre f_t est ajouté pour améliorer le modèle. La nouvelle prédiction devient :

$$\hat{y}_i(t) = \hat{y}_i(t-1) + f_t(x_i)$$

Pour déterminer l'arbre f_t , XGBoost minimise l'approximation de la fonction de perte en utilisant les gradients. La fonction objectif à minimiser pour l'étape t est :

$$L(t) = \sum_{i=1}^n L(y_i, \hat{y}_i(t-1) + f_t(x_i)) + \Omega(f_t)$$

où $\Omega(f_t)$ est un terme de régularisation qui contrôle la complexité de l'arbre f_t .

L'approximation de Taylor au second ordre de la fonction de perte $L(y_i, \hat{y}_i)$ autour de la prédiction actuelle $\hat{y}_i(t-1)$ est donnée par :

$$L(y_i, \hat{y}_i) \approx L(y_i, \hat{y}_i(t-1)) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)$$

où :

$$g_i = \left. \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i} \right|_{\hat{y}_i = \hat{y}_i(t-1)}$$

et

$$h_i = \left. \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \right|_{\hat{y}_i = \hat{y}_i(t-1)}$$

on aura donc :

$$L(t) \approx \sum_{i=1}^n \left[L(y_i, \hat{y}_i(t-1)) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Pour minimiser cette fonction objective, XGBoost construit l'arbre de décision f_t en divisant les données de manière à minimiser la perte :

$$L(t) \approx \sum_j \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T$$

avec :

- $G_j = \sum_{i \in I_j} g_i$
- $H_j = \sum_{i \in I_j} h_i$
- I_j : l'ensemble des indices des observations dans la feuille j
- w_j : la valeur prédictive associée à la feuille j .
- λ : Un terme de régularisation ajouté sur les poids w_j
- γ : Un terme de pénalité ajouté pour chaque feuille T .
- T : le nombre de feuilles.

La solution pour w_j est :

$$w_j = -\frac{G_j}{H_j + \lambda}$$

Après la construction de l'arbre, les prédictions sont mises à jour :

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

la figure ci dessous résume le fonctionnement de l'algorithme xgboost

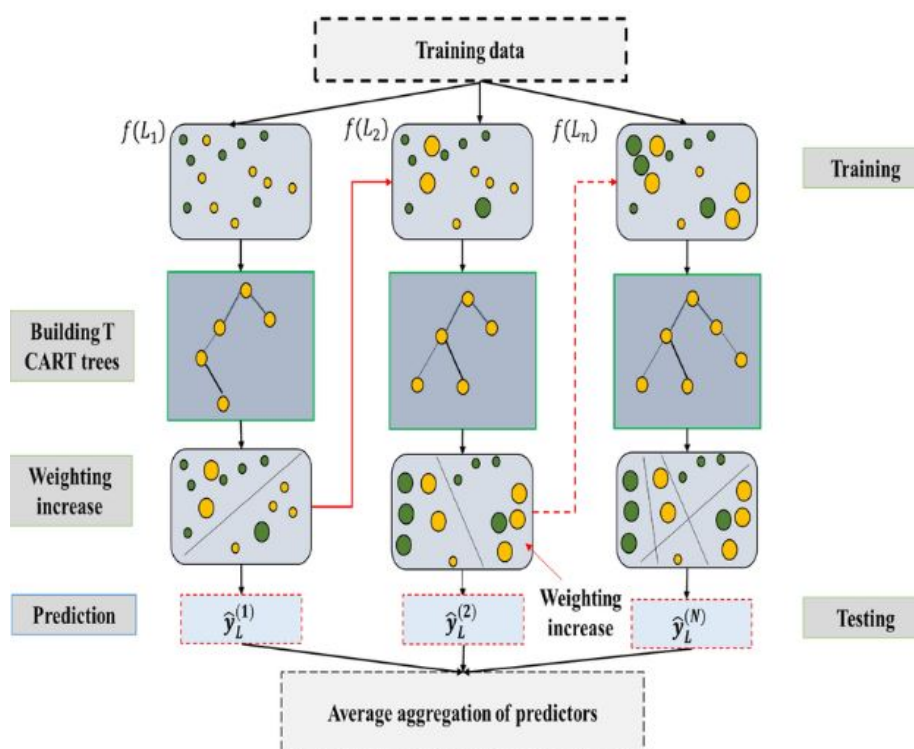


Figure 48: illustration du fonctionnement de XGBOOST

III. Application de l'algorithme XGBOOST

Avant d'appliquer cet algorithme, il est primordial de rendre les variables catégorielles binaires $\{0, 1\}$. En effet, la majorité des algorithmes d'apprentissage automatique ne sont pas en mesure de traiter les variables

catégorielles de manière indépendante. Ils exigent que les informations soient enregistrées en format numérique. Les variables catégorielles sont converties en vecteurs de zéros et de uns par le one-hot encoding, ce qui rend compatibles avec ces algorithmes .

III.1. Modélisation de la fréquence

III.1.1. Recherche du meilleur nombre d'itérations

XGBoost est un algorithme de boosting, c'est-à-dire qu'il construit le modèle de manière progressive. À chaque itération, il ajoute un nouvel arbre de décision qui vise à corriger les erreurs des arbres précédents, améliorant ainsi progressivement la performance globale du modèle.

A ce propos il faut fixer le nombre d'itérations (nrounds) dans le modèle construit, Par ailleurs, Une validation croisée est réalisée afin de déterminer le nombre d'itérations (nrounds) dans le modèle construit, en utilisant un découpage en 5 partitions. Les RMSE en fonction des différentes valeurs du nombre d'itérations sont présentées dans le graphique ci-dessous:

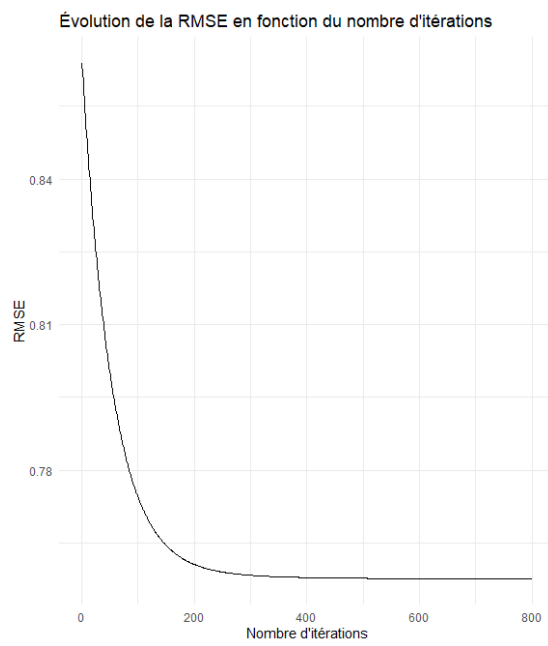


Figure 49: Evolution de la RMSE sur la base d'apprentissage en fonction du nombre d'itérations

nous pouvons voir que la RMSE se stabilise après 500 itérations. Nous avons décidé de retenir un modèle avec `nrounds= 500` .

III.1.2.Importance des variables

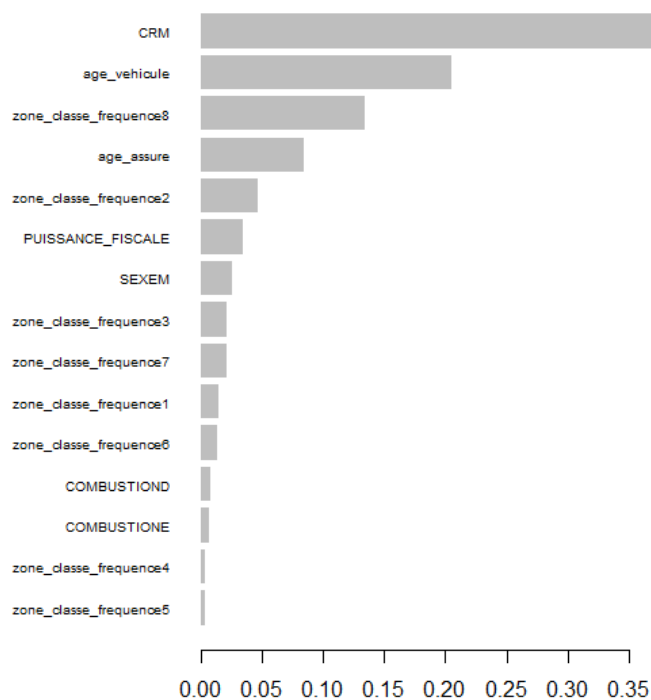


Figure 50: Importance des variables dans la modélisation de la fréquence

Le graphique d'importance des variables montre que CRM et âge du véhicule ont les scores les plus élevés.

Nous pouvons conclure que ces variables sont les prédominantes dans la construction du modèle XGBoost pour la prédiction de la fréquence des sinistres.

III.2. Modélisation du coût moyen

La même approche sera utilisée pour la modélisation du coût moyen.

III.2.1. Recherche du meilleur nombre d'itérations

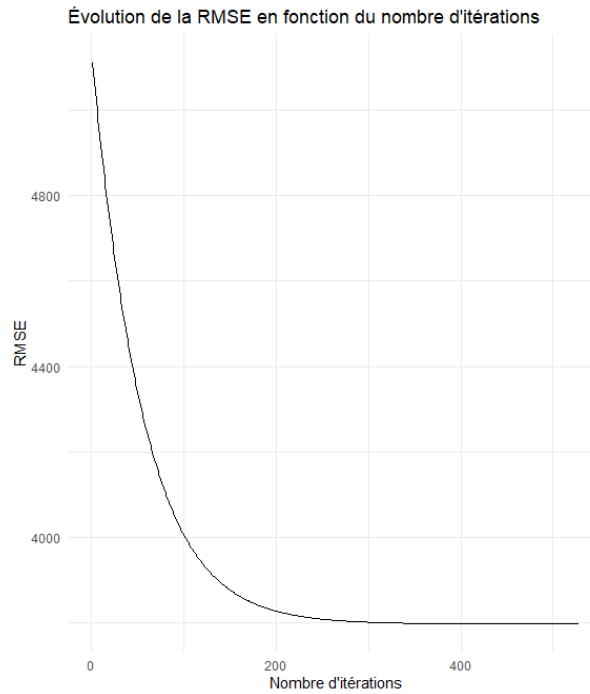


Figure 51: Evolution de la RMSE sur la base d'apprentissage en fonction du nombre d'itérations

d'après le graphique de l'évolution de la RMSE en fonction du nombre d'itérations, nous pouvons justifier le choix d'un nrounds de 400

III.2.2. Importance des variables

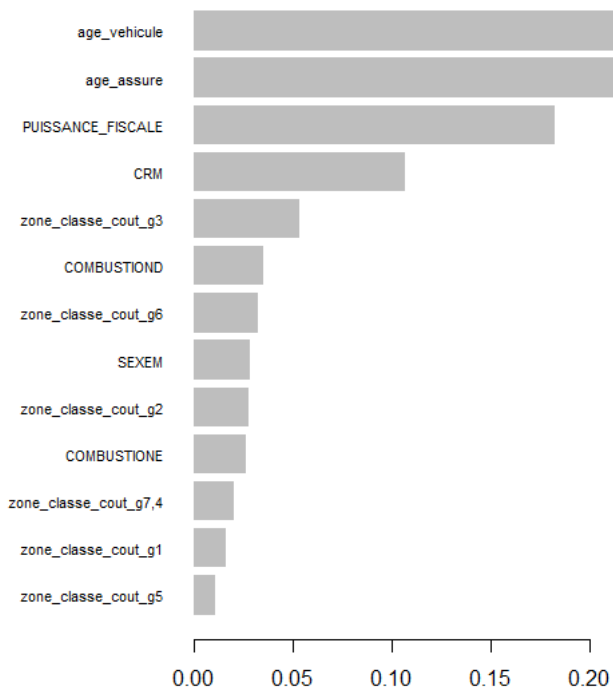


Figure 52: Importance des variables dans la modélisation du coût moyen

Nous remarquons un changement significatif dans l'importance des variables entre la prédiction de la fréquence des sinistres et celle du coût moyen. Alors que le CRM était le plus influent pour la fréquence, l'âge du véhicule devient prédominant pour le coût moyen.

L'âge du conducteur et la puissance fiscale restent également des variables importantes pour prédire le coût moyen des sinistres, mais le CRM perd de son poids et se retrouve en quatrième position.

IV. Application de l'algorithme CART

IV.1. Modélisation de la fréquence

IV.1.1. Arbre maximal

Comme mentionné dans le chapitre sur la construction du zonier tarifaire, la première étape consiste à élaborer un arbre maximal. Pour ce faire, le paramètre de complexité cp est défini à zéro.

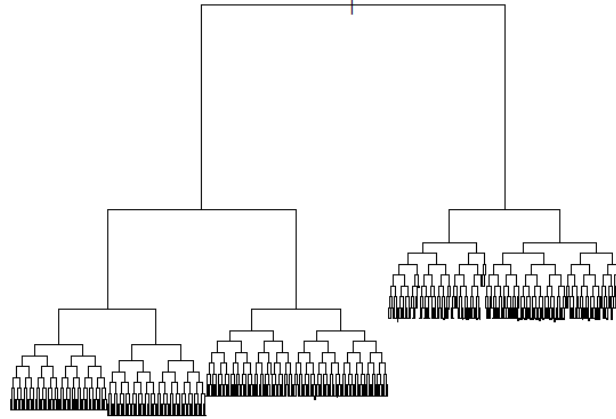


Figure 53: l'arbre maximal de la fréquence

L'arbre maximal divise trop le jeu de données. Donc, on procède à un élagage pour avoir une segmentation optimale.

IV.1.2. Elagage de l'arbre

De même, nous allons choisir le paramètre de complexité qui minimise l'erreur de la validation croisée.

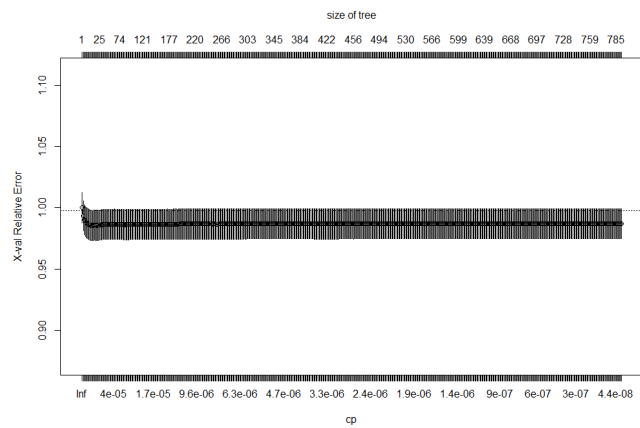


Figure 54: Graphique de l'erreur de la validation croisée

	CP	nsplit	rel error	xerror	xstd
11	1.140071e-04	11	0.9853419	0.9855185	0.01230593
12	9.823765e-05	12	0.9852279	0.9853830	0.01230476
13	9.465262e-05	13	0.9851296	0.9853365	0.01230368
14	9.271574e-05	14	0.9850350	0.9853838	0.01230295
15	8.887067e-05	15	0.9849423	0.9853346	0.01230233
16	8.344393e-05	22	0.9843202	0.9853111	0.01230213
17	6.766628e-05	23	0.9842367	0.9854269	0.01229904
18	6.429932e-05	24	0.9841690	0.9858262	0.01229935
19	6.412063e-05	30	0.9837832	0.9858447	0.01229918
20	5.960850e-05	31	0.9837191	0.9858408	0.01229886
21	5.613129e-05	33	0.9835999	0.9860254	0.01229900

Figure 55: La sortie cptable de l'arbre maximal

Nous remarquons que la valeur minimale de **xerror** est de **0,9853111**, donnant ainsi un arbre optimal avec **23** feuilles .

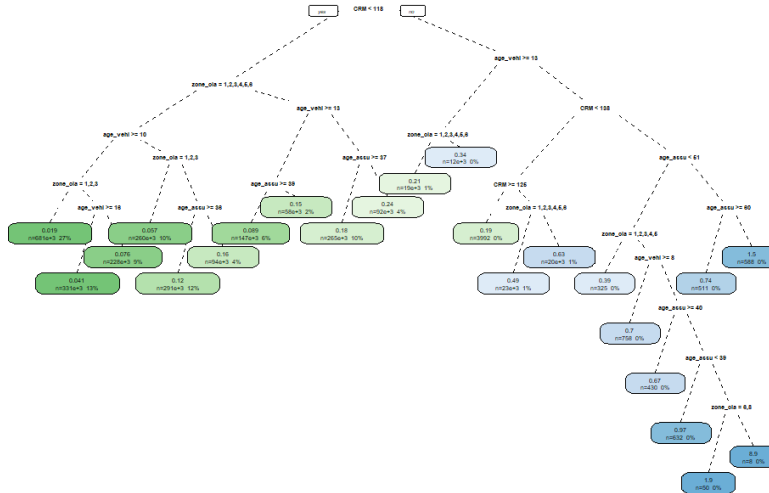


Figure 56: L'arbre optimal de la fréquence

L'arbre optimal possède 23 feuilles terminales (la variable à expliquer Y présente 23 cas de fréquence de sinistralité). Chaque observation fait partie d'une feuille en fonction de ses variables explicatives et est liée à une valeur de fréquence de sinistralité.

IV.2.Modélisation du coût moyen

IV.2.1.Arbre maximal

De même que pour la modélisation de la fréquence pour le modèle CART, nous construisons l'arbre maximal en ne pénalisant pas la complexité de l'arbre (cp=0).

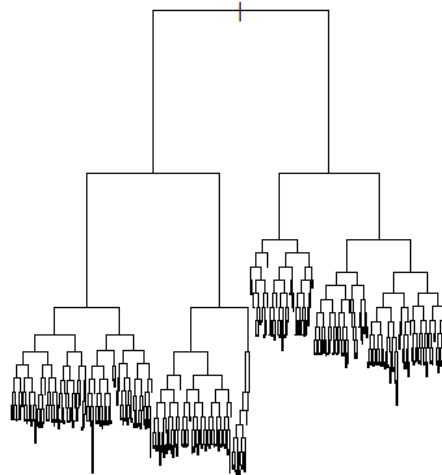


Figure 57: L'arbre maximal du coût moyen

De même, l'arbre maximal doit être élagué pour obtenir la segmentation optimale.

IV.2.2. Elagage de l'arbre

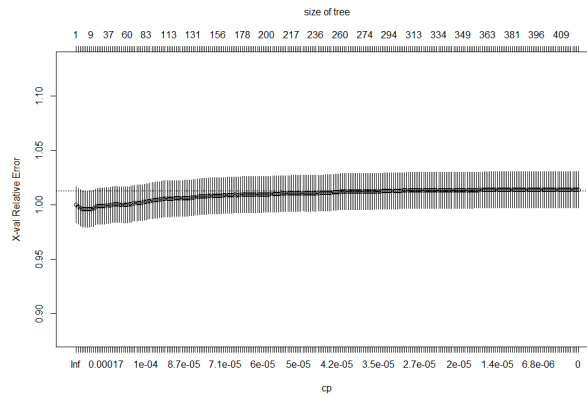


Figure 58: Graphique de l'erreur de la validation croisée

	CP	nsplit	rel error	xerror	xstd
1	0.0018796891	0	1.0000000	1.0000348	0.01654036
2	0.0015404450	1	0.9981203	0.9986411	0.01655472
3	0.0007545556	2	0.9965799	0.9972635	0.01655200
4	0.0005024654	3	0.9958253	0.9961911	0.01655317
5	0.0004662644	4	0.9953228	0.9963163	0.01655427
6	0.0003748723	5	0.9948566	0.9961643	0.01655383
7	0.0003175939	6	0.9944817	0.9961707	0.01655410
8	0.0002790601	7	0.9941641	0.9963060	0.01655653
9	0.0002447576	8	0.9938851	0.9963767	0.01654740
10	0.0002162500	9	0.9936403	0.9966545	0.01654219
11	0.0001830442	10	0.9934240	0.9980062	0.01655064

Figure 59: La sortie cptable de l'arbre maximal

En suivant la même approche que précédemment, nous allons choisir pour l'élagage de l'arbre optimal le cp tel que l'erreur de validation croisée soit minimale c'est à dire **0,9961643**.

Sur le graphique, l'arbre optimal ne comprend que 6 feuilles terminales.

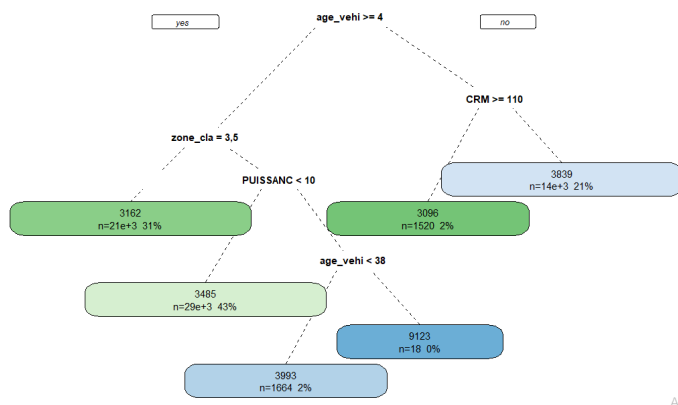


Figure 60: l'arbre optimal du coût moyen

Les feuilles terminales sont peu nombreuses, ce qui signifie que les prédictions des coûts moyens ne peuvent prendre que 6 valeurs .

On peut expliquer ce nombre de feuilles par l'homogénéité de nos données par rapport au coût moyen.

V.Conclusion

Dans ce chapitre, nous avons examiné en détail l'utilisation des techniques d'apprentissage automatique, CART et XGBoost, dans le cadre de la modélisation de la fréquence et du coût moyen des sinistres. Ainsi, nous avons commencé par une explication approfondie du cadre théorique de l'algorithme XG-BOOST , soulignant les principes essentiels et les concepts essentiels indispensables à sa compréhension avant la mise en œuvre des différentes étapes de construction des deux modèles,

Chapitre 6: Comparaison des différents modèles

I.Introduction

Comme vu dans les chapitres précédents, nous avons élaboré quatre modèles de tarification.

Le premier utilise le GLM avec une segmentation obtenue par CART, le deuxième utilise le GLM avec la segmentation proposée par l'organisme, le troisième utilise l'algorithme CART, et le dernier utilise l'algorithme XG-Boost.

Dans ce chapitre, nous allons comparer les résultats des différents modèles en nous basant sur le calcul de la RMSE. De plus, nous comparerons les différents tarifs obtenus et, enfin, nous élaborerons une application VBA permettant de calculer la prime pure.

II.Comparaison de l'Erreur Quadratique Moyenne

L'erreur quadratique moyenne (RMSE) est une mesure couramment utilisée pour mesurer la précision d'un modèle de prévision. Il s'agit d'une mesure statistique qui calcule l'ampleur moyenne des différences entre les valeurs prédites et observées. RMSE fournit une valeur unique qui représente les performances globales du modèle, permettant une comparaison facile entre différents modèles ou techniques de prévision.

Rappelons la formule de calcul :

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

avec:

- n : Nombre d'observations
- \hat{y}_i : Valeur prédite pour l'observation i
- y_i : Valeur observée pour l'observation i

-Application :

Pour comparer les différents modèles, nous allons calculer le RMSE sur la base de test. Les tableaux ci-dessous résumant les résultats obtenus pour la responsabilité civile matérielle et corporelle.

modèle	fréquence	coût moyen
GLM1	0.8057	2977.66
GLM2	0.8030	4076.24
CART	0.7612	3655.45
XG-Boost	0.7633	3872.02

Table 4: Synthèse des RMSE des modèles sur la base test (segment matériel)

Nous remarquons que pour la modélisation de la fréquence des sinistres matérielle, le modèle dont l'erreur est la plus faible (RMSE= 0.7612) est bien CART, il est suivi de XG-Boost avec une valeur de RMSE égale à 0.7633.

Quant à la modélisation du coût du sinistre, c'est le modèle GLM1 qui l'emporte avec une valeur de RMSE égale à 2977.66, il est suivi de CART dont la RMSE vaut 3655.45.

modèle	fréquence	coût moyen
GLM1	0.1649	27271.31
GLM2	0.1421	26423.21
CART	0.1370	18986.12
XG-Boost	0.1277	22494.74

Table 5: Synthèse des RMSE des modèles sur la base test (segment corporel)

Il est observé que pour la modélisation de la fréquence des sinistres corporels, le modèle avec la plus faible erreur (RMSE = 0.1277) est XG-Boost, suivi de CART avec une valeur de RMSE égale à 0.1370.

En ce qui concerne la modélisation du coût du sinistre, le modèle CART est le plus performant avec une valeur de RMSE de 18986.12.

III.Synthèse des primes pures

-le modèle GLM1 :

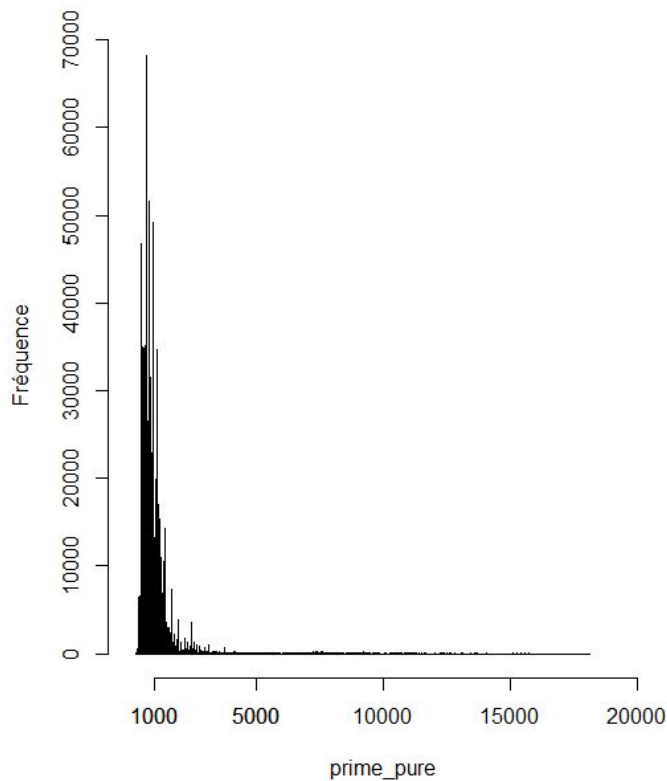


Figure 61: Distribution des primes pures pour le modèle GLM1

Nous avons en abscisse les primes pures modélisées par le modèle GLM1 et en ordonnée le nombre total d'observations y associé. Nous remarquons que les primes pures prennent généralement des valeurs inférieures à 5000, avec une forte concentration entre 800 et 1000.

Min	1er Qu.	Médiane	Moyenne	3e Qua.	Max	Ecart type
300.0863	663.8016	803.03	893.16	922.02	18137.78	536.4182

Table 6: Statistiques descriptives des prédictions de primes pures sur la base production

Selon les statistiques, les primes varient considérablement, allant de 300.0863 à 18137.78. La tendance médiane de 803.03 et la moyenne de 893.16 mettent en évidence une tendance centrale située entre 800 et 900.

–le modèle GLM2 :

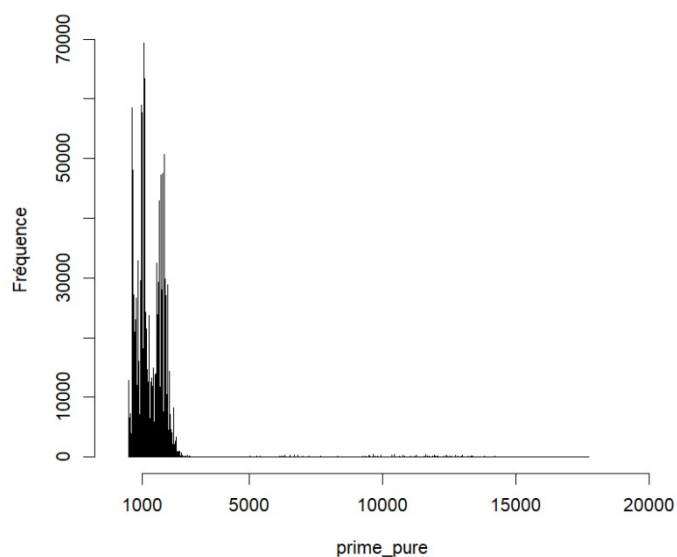


Figure 62: Distribution des primes pures pour le modèle GLM2

Min	1er Qu.	Médiane	Moyenne	3e Qua.	Max	Ecart type
474.1738	955.9475	1137.91	1288.19	1631.14	17714.54	759.4591

Table 7: Statistiques descriptives des prédictions de primes pures sur la base production

Selon les statistiques, les primes varient considérablement, allant de 474.1738 à 17714.54. La tendance médiane de 1137.91 et la moyenne de 1288.19 indiquent une tendance centrale située entre 1100 et 1300.

-le modèle CART :

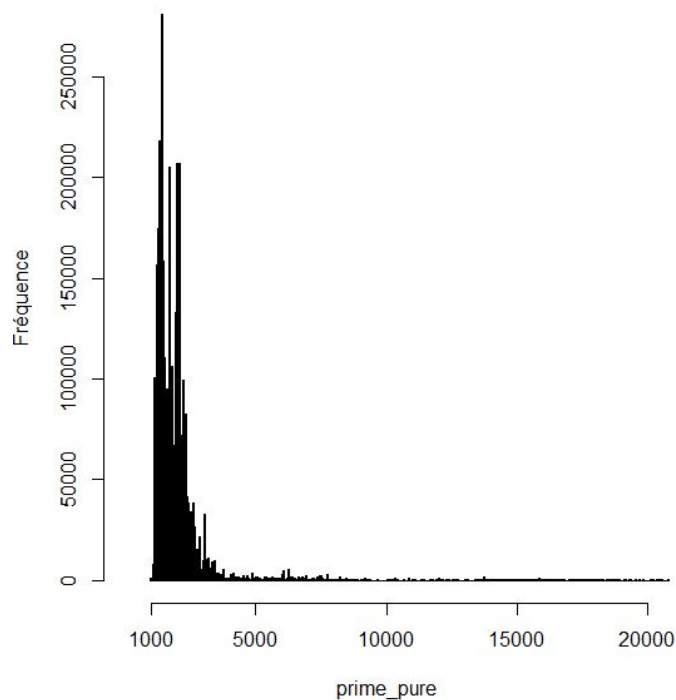


Figure 63: Distribution des primes pures pour le modèle CART

Min	1er Qu.	Médiane	Moyenne	3e Qua.	Max	Ecart type
791.25	1345.443	1678.72	2028.12	2197.82	3234657	5100.144

Table 8: Statistiques descriptives des prédictions de primes pures sur la base production

Les données mettent en lumière une variété de primes, allant de 791.25 à 3234657. La médiane étant de 1678.72 et la moyenne de 2028.12, les primes sont concentrées entre 1700 et 2000. Toutefois, l'écart type élevé de 5100.14 met en évidence une grande dispersion des données par rapport à la moyenne, ce qui suggère une grande variabilité dans les primes.

-le modèle XG-Boost :

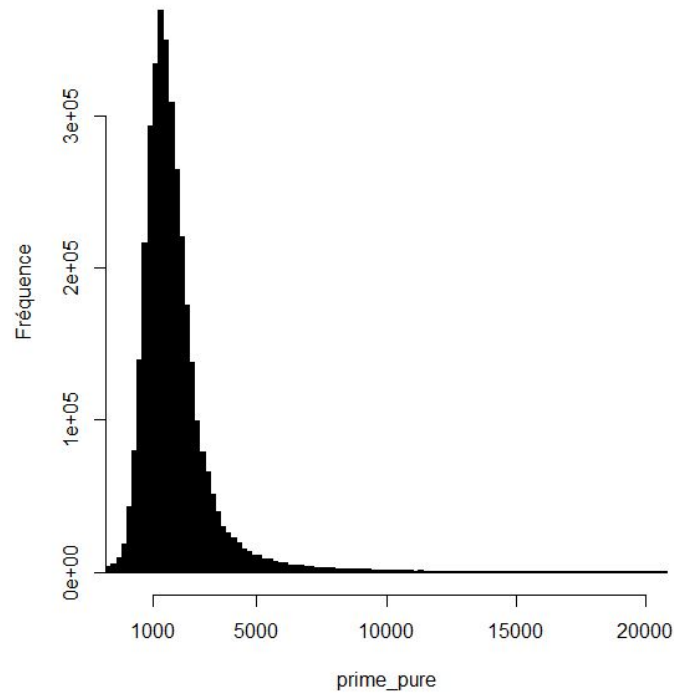


Figure 64: Distribution des primes pures pour le modèle XG-Boost

Min	1er Qu.	Médiane	Moyenne	3e Qua.	Max	Ecart type
-1463482	1045.049	1536.086	2013.76	2203.85	22226850	18770.75

Table 9: Statistiques descriptives des prédictions de primes pures sur la base production

Les données sur les primes montrent une grande variabilité, allant de -1463482 à 22226850. La tendance centrale, avec une médiane de 1536.086 et une moyenne de 2013.76, se situe entre 1500 et 2000. Cependant, l'écart type élevé de 18770,75 met en évidence une grande dispersion par rapport à la moyenne.

IV.construction d'une application VBA Excel

Dans cette partie, nous allons créer une application VBA Excel qui permet de calculer la prime pure et la prime chargée(une charge de 30%) à l'aide des différents modèles en saisissant les caractéristiques du conducteur.

NB: Il convient de souligner que, étant donné la complexité du modèle XG-Boost, il a été compliqué de le charger dans VBA Excel. Ainsi, nous restreindrons nos choix aux trois autres modèles.

pour comparer nos algorithmes, ci dessous un exemple de calcul de la prime pour un conducteur.

Modèle	Prime pure	Prime Chargee
CART	2096	2725
GLM1	1500	1951
GLM2	1321	1718

Figure 65: interface de l'application VBA

Nous remarquons que la prime pure diffère d'un modèle à un autre. En effet, pour le modèle GLM1, la prime pure est de 1500 ; pour le modèle GLM2, elle est de 1321, et enfin, pour le modèle CART, elle est de 2096. Cette différence entre les primes s'explique par la méthode de segmentation des variables ainsi que par l'algorithme de prédiction de chaque modèle.

V.Conclusion

Dans ce chapitre, nous avons effectué une comparaison des différents modèles de prédiction en nous basant sur le calcul de l'erreur quadratique moyenne. Ensuite, nous avons synthétisé les primes des différents modèles en calculant les principales statistiques descriptives. Enfin, nous avons

développé une application VBA Excel permettant de calculer la prime en saisissant simplement les données du conducteur.

Conclusion générale

La tarification des produits d'assurance revêt une importance capitale pour les compagnies d'assurance, particulièrement dans le secteur automobile, qui est hautement concurrentiel et constitue la principale source de revenus dans le domaine de l'assurance des biens et de la responsabilité. Les assureurs sont ainsi confrontés au défi de concevoir des tarifications justes et précises, tout en préservant le principe fondamental de la mutualisation des risques, afin de mieux appréhender les risques inhérents à leur activité. Pour attirer de nouveaux clients tout en maintenant leur rentabilité, les assureurs doivent être capables de segmenter efficacement leurs risques.

Dans cette perspective, ce mémoire vise à explorer différentes méthodes de tarification en modélisant deux aspects essentiels : la fréquence des sinistres et le coût moyen au sein d'un portefeuille d'assurance responsabilité civile automobile. Traditionnellement, les compagnies d'assurance utilisent des méthodes économétriques classiques, telles que les modèles linéaires généralisés (GLM), pour établir leurs tarifs. Cependant, l'émergence de nouveaux algorithmes innovants, en partie stimulée par l'accessibilité croissante aux données, a ouvert de nouvelles perspectives. Ces algorithmes, regroupés sous le terme d'apprentissage statistique (machine learning), visent soit à prédire des valeurs, soit à classer des individus, offrant ainsi de nouvelles approches pour la tarification des risques en assurance.

Nous avons alors réalisé une étude comparative des performances réalisées par les modèles traditionnels (GLM) qui nous serviront de modèle de base, et de quelques modèles de machine learning : les arbres de décision CART, et l'eXtreme Gradient Boosting Machine. L'étude menée montre que les approches data science sont pertinentes et offrent des performances comparables à celles des GLM.

Bibliographie et Webographie

- [1] R. Bellina, *Méthodes d'apprentissage appliquées à la tarification non-vie*, Mémoire Institut des Actuares, 2014.
- [2] H.BENABDERRAHMAN et S.EL KHALIFA . *Elaboration d'un Zonier automobile Modélisation de risque et étude d'impact* ,2017
- [3] J.DELIMADJEON, *Modélisation statistique à la création d'un zonier en tarification automobile* , Mémoire Institut des Actuares,2021
- [4] E.BOUTAHAR ,*Application à la tarification automobile de méthodes de partitionnement récursif de modèles linéaires généralisés*,Mémoire Institut des Actuares
- [5] F.Zouggagh ,*Tarification automobile à l'aide de modèles de machine learning et apport des données télématiques*,Mémoire Institut des Actuares,2018
- [6] F.BOUTTIER,*Modélisation de la prime pure de la garantie rc automobile*,Mémoire Institut des Actuares,2015
- [7] Y.LUO,*Amélioration de la modélisation de sinistres graves à l'aide d'une approche d'apprentissage*,Mémoire Institut des Actuares,2015
- [8] K.MEZIANI,*Méthodes pour les modèles de régression*,Cours de Master 2 Mathématiques Appliquées à l'Université Paris Dauphine,2019
- [9] C.CHESNEAU,*Introduction aux arbres de decision(de type CART)*,Université Caen-Normandie,2023
- [10] F.SANTOS,*Arbres de décision*,Université de Bordeaux,2015
- [11] V.GRARI,*Impact des données exogènes sur la tarification en santé.*,Mémoire Institut des Actuares,2015
- [12] A.GUILLOT,*Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?* ,Mémoire Institut des Actuares,2015
- [13] F.LAGADEC,*Tarification d'un contrat de complémentaire santé par un modèle linéaire généralisé* ,Mémoire Institut des Actuares,2009
- [14] F. FAGHRI,*Random Forests& XGBoost*,University of Toronto,2019

- <https://datacorner.fr/xgboost/>
- <https://datafuture.fr/post/faire-tourner-xgboost-sous-r/>
- https://rpubs.com/mdhafer/arbres_decision
- <https://www.acaps.ma/fr>

- <https://www.hcp.ma/>
- <https://www.actuarialab.net/>

ANNEXES

–Analyse graphique des variables :

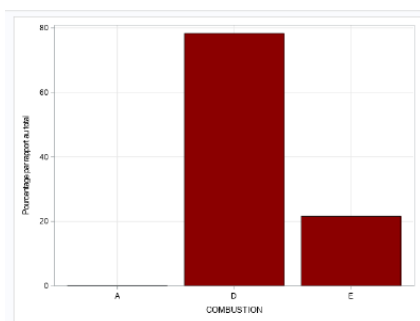


Figure 66: Distribution de combustion

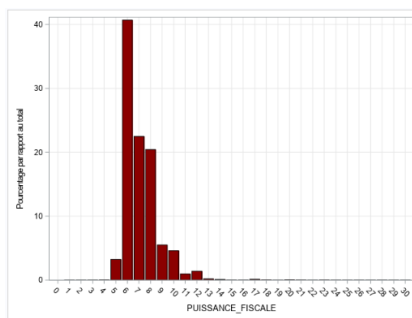


Figure 67: Distribution de la puissance fiscale

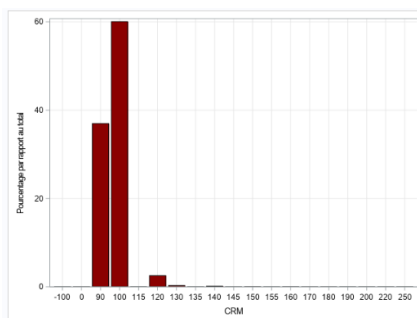


Figure 68: Distribution de CRM

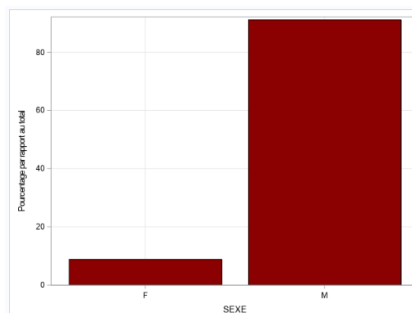


Figure 69: Distribution de sexe

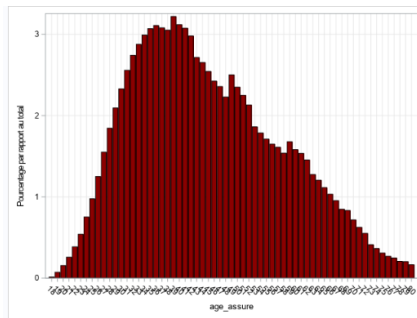


Figure 70: Distribution de l'âge du conducteur

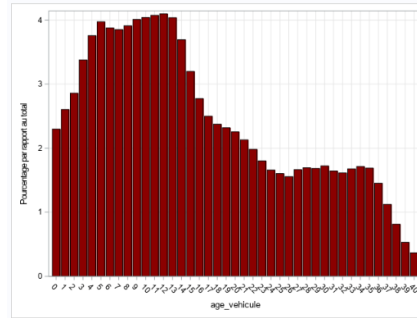


Figure 71: Distribution de l'âge du véhicule

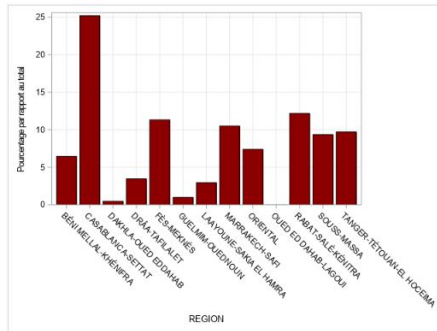


Figure 72: Distribution des régions

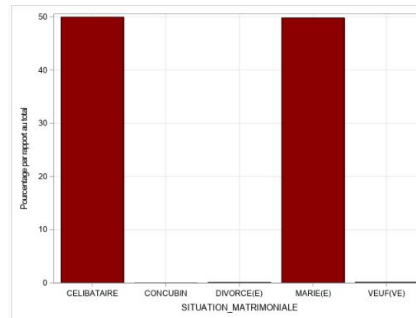


Figure 73: Distribution de la situation matrimoniale

– Zonier obtenu par CART pour la fréquence :

Classes	Caractéristiques
classe 1	Densité < 13045, taux d'activité >= 49%, prct_travaillant_domicile < 1.6%
classe 2	Densité < 4049, taux d'activité < 49%, prct_depla_voiture < 33%
classe 3	Densité ∈ [4049, 13045[, taux d'activité < 49%, prct_deplac_voiture < 33%
classe 4	Densité < 13045, taux d'activité >= 49%, prct_travaillant_domicile ∈ [1.6%, 2.5%[
classe 5	Densité < 13045, taux d'activité < 49%
classe 6	Densité < 13045, taux d'activité >= 49%, prct_travaillant_domicile ∈ [2.5%, 4.7%[
classe 7	Densité < 13045, taux d'activité >= 49%, prct_travaillant_domicile >= 4.7%
classe 8	Densité >= 13045

Table 10: Zonier obtenu par CART pour la fréquence

– Zonier obtenu par CART pour le coût moyen :

Classes	Caractéristiques
classe 1	taux d'activité $\geq 47\%$, prct_travaillant_domicile $\geq 2.9\%$,prct_Ind_pendants $< 22\%$
classe 2	taux d'activité $\geq 47\%$, prct_travaillant_domicile $< 2.9\%$, Densité $\geq 6915\%$
classe 3	taux d'activité $\geq 47\%$, prct_travaillant_domicile $\geq 2.9\%$, prct_Ind_pendants $\geq 22\%$
classe 4	taux d'activité $< 47\%$,Taux_de_chomage $\geq 19\%$
classe 5	taux d'activité $\geq 47\%$, prct_travaillant_domicile $< 2.9\%$, Densité < 6915
classe 6	taux d'activité $< 47\%$, Taux_de_chomage $< 15\%$
classe 7	taux d'activité $< 47\%$, Taux_de_chomage $\in [15\%,19\%[$

Table 11: Zonier obtenu par CART pour le coût moyen

– segmentation proposée par l’organisme (utilisée dans le modèle GLM2):

Variable	Condition	Classe
Puissance fiscale sans croisement	puissance_fiscale < 5	classe1
	puissance_fiscale = 5 ou puissance_fiscale = 6	classe2
	puissance_fiscale = 7	classe3
	puissance_fiscale = 8 ou puissance_fiscale = 9	classe4
	puissance_fiscale = 10	classe5
	puissance_fiscale >= 11	classe6
Puissance fiscale avec croisement	nv_combustion = "D" et puissance_fiscale <= 4	classe1
	nv_combustion = "D" et puissance_fiscale = 5	classe2
	nv_combustion = "D" et puissance_fiscale = 7 ou 6	classe3
	nv_combustion = "D" et puissance_fiscale >= 8	classe4
	nv_combustion = "A et E" et puissance_fiscale <= 6	classe1
	nv_combustion = "A et E" et puissance_fiscale = 8 ou 7	classe2
	nv_combustion = "A et E" et puissance_fiscale = 10 ou	classe3
nv_combustion = "A et E" et puissance_fiscale > 10	classe4	
Age assure	age_assure < 25	classe1
	25 <= age_assure < 40	classe2
	40 <= age_assure < 60	classe3
	age_assure >= 60	classe4
Age vehicule	age_vehicule < 2	classe1
	2 <= age_vehicule < 5	classe2
	age_vehicule >= 5	classe3
CRM	CRM <= 90	classe1
	CRM > 90 et CRM < 125	classe2
	CRM >= 125 et CRM < 175	classe3
	CRM >= 175	classe4
ZONE	Casablanca	classe1
	Rabat	classe2
	zone faible	classe3
	zone moyenne	classe4
	zone forte	classe5

Figure 74: segmentation proposée par l’organisme

– modélisation du coût moyen matériel (GLM2):

Obs	Parameter	Level1	DF	Estimate	StdErr	LowerWaldCL	UpperWaldCL	ChiSq	ProbChiSq
1	Intercept		1	7.7462	0.0135	7.7197	7.7726	329441	<.0001
2	SEXE	F	1	-0.0438	0.0080	-0.0593	-0.0282	30.26	<.0001
3	SEXE	M	0	0.0000	0.0000	0.0000	0.0000	.	.
4	nv_zone_pdv5	FAIBLE	1	0.0952	0.0075	0.0804	0.1099	159.75	<.0001
5	nv_zone_pdv5	MOYENNE,FORTE,RABAT	1	0.1081	0.0068	0.0947	0.1215	250.71	<.0001
6	nv_zone_pdv5	CASABLANCA	0	0.0000	0.0000	0.0000	0.0000	.	.
7	nv_pf9	classe1-2	1	-0.1747	0.0127	-0.1997	-0.1497	187.80	<.0001
8	nv_pf9	classe3	1	-0.1168	0.0142	-0.1447	-0.0889	67.26	<.0001
9	nv_pf9	classe4	1	-0.1100	0.0134	-0.1362	-0.0838	67.75	<.0001
10	nv_pf9	classe5-6	0	0.0000	0.0000	0.0000	0.0000	.	.
11	nv_Combustion	A et E	1	-0.0284	0.0081	-0.0443	-0.0126	12.42	0.0004
12	nv_Combustion	D	0	0.0000	0.0000	0.0000	0.0000	.	.
13	nv_age_assure1	classe1	1	0.1005	0.0204	0.0606	0.1404	24.33	<.0001
14	nv_age_assure1	classe2	1	-0.0207	0.0062	-0.0329	-0.0086	11.20	0.0008
15	nv_age_assure1	classe3-4	0	0.0000	0.0000	0.0000	0.0000	.	.
16	age_vehicule_classe	age_vehicule_classe1	1	-0.0699	0.0100	-0.0895	-0.0503	48.96	<.0001
17	age_vehicule_classe	age_vehicule_classe2	1	-0.0157	0.0079	-0.0312	-0.0002	3.96	0.0466
18	age_vehicule_classe	age_vehicule_classe3	0	0.0000	0.0000	0.0000	0.0000	.	.
19	nv_CRM1	classe2	1	0.0559	0.0062	0.0437	0.0680	81.24	<.0001
20	nv_CRM1	classe3-4	1	0.0705	0.0255	0.0206	0.1205	7.66	0.0056
21	nv_CRM1	classe1	0	0.0000	0.0000	0.0000	0.0000	.	.
22	Scale		1	1.1848	0.0021	1.1808	1.1889	–	–

Obs	Criterion	DF	Value	ValueDF
1	Deviance	17E4	232698.6456	1.4039
2	Scaled Deviance	17E4	165761.0000	1.0001
3	Pearson Chi-Square	17E4	232698.6456	1.4039
4	Scaled Pearson X2	17E4	165761.0000	1.0001
5	Log Likelihood	–	-263317.5085	–
6	Full Log Likelihood	–	-263317.5085	–
7	AIC (smaller is better)	–	526665.0171	–
8	AICC (smaller is better)	–	526665.0200	–
9	BIC (smaller is better)	–	526815.2916	–

– modélisation de la fréquence matérielle (GLM2):

Obs	Parameter	Level1	Level2	DF	Estimate	StdErr	LowerWaldCL
1	Intercept			1	-0.1804	0.0209	-0.2214
2	SEXE	F		1	0.2725	0.0062	0.2602
3	SEXE	M		0	0.0000	0.0000	0.0000
4	Zone_PDV	CASABLANCA		1	0.2263	0.0071	0.2123
5	Zone_PDV	FAIBLE		1	-0.6028	0.0074	-0.6172
6	Zone_PDV	FORTE		1	-0.8698	0.0095	-0.8884
7	Zone_PDV	MOYENNE		1	-0.4895	0.0082	-0.5055
8	Zone_PDV	RABAT		0	0.0000	0.0000	0.0000
9	nv_CRM1	classe1		1	-1.2936	0.0189	-1.3307
10	nv_CRM1	classe2		1	-0.9725	0.0188	-1.0094
11	nv_CRM1	classe3-4		0	0.0000	0.0000	0.0000
12	nv_Combustion*nv_pf4	A et E	classe1	1	-0.3241	0.0096	-0.3428
13	nv_Combustion*nv_pf4	A et E	classe2	1	-0.2992	0.0097	-0.3182
14	nv_Combustion*nv_pf4	A et E	classe3-4	1	-0.1773	0.0135	-0.2037
15	nv_Combustion*nv_pf4	D	classe1-2	1	0.0491	0.0132	0.0232
16	nv_Combustion*nv_pf4	D	classe3	1	-0.0912	0.0054	-0.1019
17	nv_Combustion*nv_pf4	D	classe4	0	0.0000	0.0000	0.0000
18	age_assure_classe	age_assure_classe1		1	0.5731	0.0164	0.5408
19	age_assure_classe	age_assure_classe2		1	0.3610	0.0068	0.3477
20	age_assure_classe	age_assure_classe3		1	0.0990	0.0065	0.0862
21	age_assure_classe	age_assure_classe4		0	0.0000	0.0000	0.0000
22	age_vehicule_classe	age_vehicule_classe1		1	0.5090	0.0078	0.4937

Obs	UpperWaldCL	ChiSq	ProbChiSq
1	-0.1395	74.46	<.0001
2	0.2847	1904.80	<.0001
3	0.0000	.	.
4	0.2402	1016.73	<.0001
5	-0.5884	6720.97	<.0001
6	-0.8511	8334.56	<.0001
7	-0.4735	3597.35	<.0001
8	0.0000	.	.
9	-1.2564	4665.15	<.0001
10	-0.9355	2663.65	<.0001
11	0.0000	.	.
12	-0.3054	1151.73	<.0001
13	-0.2803	959.98	<.0001
14	-0.1508	172.89	<.0001
15	0.0750	13.77	0.0002
16	-0.0806	282.29	<.0001
17	0.0000	.	.
18	0.6053	1215.14	<.0001
19	0.3743	2828.57	<.0001
20	0.1118	230.76	<.0001
21	0.0000	.	.
22	0.5243	4243.76	<.0001

Obs	Criterion	DF	Value	ValueDF	pvalue
1	Pearson Chi-Square	36E5	3596833.9423	1.0002	0.38965
2	Scaled Pearson X2	36E5	3596833.9423	1.0002	0.38965
3	AIC (smaller is better)	—	1520943.3921	—	.
4	BIC (smaller is better)	—	1521192.2040	—	.