



المندوبية السامية للتخطيط
HAUT-COMMISSARIAT AU PLAN



Institut National de Statistique
et d'Economie Appliquée

Projet de Fin d'Etudes

Modélisation de la rétention client en assurance automobile

Préparé par : **EL ONSRI Anas**
EL-ADAMI Hicham

Sous la direction de : **M. SAID Khalil (INSEA)**
Mme. SAID Salma (AXA Assurance Maroc)
M. ALOUAN Imad Eddine (AXA Assurance Maroc)

*Soutenu publiquement comme exigence partielle en vue de
l'obtention du Diplôme d'Ingénieur d'Etat*

Option : Actuariat-Finance

Devant le jury composé de :

M. SAID Khalil
M. EL ORAYBI Amal
Mme. SAID Salma
M. ALOUAN Imad Eddine

Septembre 2020/ PFE N°9

Résumé

La concurrence en assurance automobile n'a cessé de s'amplifier au cours des dernières années, mettant les acteurs du marché devant la nécessité de trouver des stratégies de croissance, capables de garantir un développement rentable de leurs portefeuilles.

Or la méthode de tarification traditionnelle, basée essentiellement sur une segmentation du portefeuille en fonction du risque de sinistralité des assurés, ne suffit plus pour relever ce défi. En effet, cette segmentation technique, aussi fine qu'elle soit, optimise la prise de risque de l'assureur, mais pas son profit.

Pour tirer un maximum de valeur de son portefeuille, l'assureur doit mieux appréhender la demande d'assurance sur son marché et intégrer cette connaissance dans sa stratégie de prix.

Dans ce document nous nous intéressons à cette problématique. Nous étudions plus particulièrement comment la modélisation de la rétention client et de l'élasticité de la demande au prix peut aider l'assureur à optimiser sa politique tarifaire au renouvellement, et constituer ainsi un levier majeur d'accroissement de sa rentabilité.

Ce mémoire a été réalisé au sein d'AXA Assurance Maroc (AAM), un acteur important dans le marché marocain d'assurance automobile. Comme c'est le cas pour tout assureur, ses dirigeants sont confrontés en permanence à un arbitrage entre leur volonté de défendre leur portefeuille en maximisant la rétention client d'une part, leur impératif de rentabilité d'autre part. De plus, pour garantir l'équilibre financier de l'entreprise, ils doivent respecter des contraintes d'évolution globale de leurs ressources.

Ce document est structuré de la manière suivante :

La première partie de ce rapport contextualise notre étude. Nous aborderons dans la deuxième partie une analyse exploratoire des données mises à notre disposition. La dernière partie est consacrée à l'élaboration du modèle de rétention modélisant la probabilité de renouvellement d'un assuré et une analyse de l'élasticité-prix des clients qui constituent notre portefeuille étudié.

Mots clés : Modèle de rétention, Élasticité-prix, Assurance automobile

Dédicaces

Je dédie ce travail :

- À mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse et leur soutien.
- À toute personne qui a contribué, de près ou de loin, à l'accomplissement de ce projet.

Anas

Je dédie ce travail :

- À mes parents qui m'ont soutenu tout au long de mon cursus scolaire.
- À tous ceux qui ont contribué à la réalisation de ce travail.

Hicham

Remerciement

Nous tenons à exprimer nos sincères remerciements à M. **ALOUAN Imad Eddine** qui nous a accordé l'opportunité d'effectuer notre stage de fin d'études au sein de son équipe.

Nous exprimons aussi notre profonde reconnaissance à notre encadrante Mme. **SAID Salma** pour ses directives précieuses et ses conseils pertinents qui nous ont été d'un appui considérable dans notre stage.

Nous adressons nos vifs remerciements à notre professeur encadrant M. **SAID Khalil**, qui par sa patience, ses conseils, et sa disponibilité notre travail a pu être mené au bon port.

Nos remerciements vont également :

- Aux membres du jury d'avoir accepté d'évaluer ce projet et pour toutes leurs remarques et leurs propositions enrichissantes ;
- À tous les professeurs qui nous ont enseigné et qui par leurs compétences nous ont soutenu dans la poursuite de nos études ;
- À tous le personnel d'AXA assurance qui nous a facilité l'insertion dans le milieu du travail ;
- À tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Merci à toutes et à tous !

Table des matières

Résumé	3
Dédicaces	4
Remerciement	5
Table des matières	6
Liste des abréviations	9
Liste des tableaux	10
Liste des figures	11
Chapitre 1 : Contexte de l'étude	12
I. Présentation de l'organisme d'accueil	12
1. AXA au monde	12
2. AXA Assurance Maroc (AAM)	13
a. Présentation.....	13
b. Organigramme	14
c. Produits commercialisés.....	15
d. Historique	16
e. Chiffres clés	16
II. Assurance automobile au Maroc.....	17
1. Introduction	17
2. Tarif du produit automobile	20
3. Cycle de vie d'une police	21
4. Vision globale du marché.....	22
III. Aperçu sur la tarification commerciale (commercial pricing)	23
1. Contexte	23
2. Eléments de la tarification commerciale	23
a. Connaissance du marché	24
i. Meilleur tarif (Best price).....	24
ii. Price testing	24
b. Modélisation de la demande	25
c. Optimisation des primes (Price optimization)	27

d. Problématique.....	28
Chapitre 2 : Analyse exploratoire des données	29
I. Présentation des données manipulées.....	29
1. Base « Contrat »	29
2. Base « Conducteur »	30
3. Base « Véhicule »	30
II. Préparation de la base de données	31
1. Création de la variable cible	31
2. Introduction des variables externes à notre base de données	34
a. Prime	34
b. Ecart à la concurrence (Best price)	34
c. Autres variables.....	34
III. Analyse exploratoire des données	35
1. Taux de rétention selon le sexe	37
2. Taux de rétention selon type d'échéance	38
3. Taux de rétention selon type de garantie	39
4. Taux de rétention en fonction de la prime	40
5. Taux de rétention en fonction de l'ancienneté du contrat	40
6. Taux de rétention en fonction de l'âge du conducteur	41
7. La rétention en fonction du taux CRM	42
8. Corrélation entre les variables explicatives deux à deux	43
Chapitre 3 : Modélisation de la rétention et étude de l'élasticité-prix	44
I. Cadre théorique des Modèles Linéaires Généralisés (GLM).....	44
II. Analyse explicative : Modélisation	44
1. Formalisation générale du modèle.....	44
a. Notations.....	44
b. Composante aléatoire	45
c. Composante déterministe	45
d. Fonction de lien	45
2. Modèle Logistique	46

a. Appartenance à la famille exponentielle	46
b. Estimation de paramètres	47
c. Significativité des variables explicatives : Test de type III	48
d. Significativité des coefficients : Test de Wald	49
e. Validation du modèle : Test d'hypothèse global ($\beta = 0$)	49
d. Performance du modèle logistique.....	50
i. Matrice de confusion.....	50
ii. Courbe ROC	51
II. Modélisation de la rétention	52
1. Formulation du modèle.....	52
2. Résultats de la modélisation	53
a. Echantillonnage.....	53
b. Estimation de paramètres	54
c. Evaluation du modèle.....	58
i. Test globale de significativité	58
ii. Matrice de confusion.....	58
iii. Courbe ROC	60
II. Calcul d'élasticité-prix	64
1. Cadre théorique.....	64
a. Rappel.....	64
b. Elasticité-prix selon le modèle de rétention	64
2. Résultats de l'élasticité-prix	66
a. Elasticité-prix en fonction de la probabilité de renouvellement	67
b. Elasticité-prix en fonction de la prime	69
3. Validation empirique des résultats « Price testing ».....	70
IV. Aperçu sur l'optimisation	70
Conclusion	73
Références.....	74
Annexe I:.....	75
Annexe II:	80

Liste des abréviations

AAM	: AXA Assurance Maroc
AUC	: Area Under the Curve
CRM	: Coefficient Réduction/Majoration
GLM	: Modèle Linéaire Généralisé
IARD	: Incendies, Accidents et Risques Divers
ROC	: Receiver Operating Characteristic

Liste des tableaux

Tableau 1.1 : Les produits d'assurance commercialisés par AAM.....	15
Tableau 1.2 : Historique d'AAM.....	16
Tableau 1.3 : les garanties auto offertes par AAM	19
Tableau 1.4 : Formules auto offertes par AAM.....	20
Tableau 2.1 : Description de la base « Contrat ».....	29
Tableau 2.2 : Description de la base « Conducteur ».....	30
Tableau 2.3 : Description de la base « Véhicule ».....	30
Tableau 2.4 : Matrice de corrélations.....	43
Tableau 3.1 : Fonctions de liens classiques.....	46
Tableau 3.2 : Matrice de confusion.....	50
Tableau 3.3 : Interprétation des valeurs du critère AUC.....	52
Tableau 3.4 : Extrait du tableau des paramètres du modèle	54
Tableau 3.5 : Les paramètres estimés du modèle retenu	55
Tableau 3.6 : Estimations des rapports des cotes.....	56
Tableau 3.7 : Tests globaux de significativité du modèle	58
Tableau 3.8 : Matrice de confusion.....	60
Tableau 3.9 : Comparaison entre les algorithmes de prédiction	62
Tableau 3.10 : Matrice de confusion.....	62

Liste des figures

Figure 1.1 : Répartition géographique du groupe AXA dans le monde	12
Figure 1.2 : Organigramme d'AAM	14
Figure 1.3 : Quelques chiffres clés sur la croissance d'AAM	16
Figure 1.4 : Répartition des primes émises en assurance automobile selon type de garantie en 2018... 17	
Figure 1.5 : Décomposition de la prime commerciale	21
Figure 1.6 : Répartition des primes émises en automobile par entreprise d'assurance en 2018..... 22	
Figure 1.7 : Eléments de la tarification commerciale	23
Figure 1.8 : Demande des affaires nouvelles en fonction de la distance par rapport au meilleur tarif .. 24	
Figure 1.9 : Exemple de distributions utilisées dans le price testing	25
Figure 1.10 : Calcul de l'élasticité-prix empirique	26
Figure 1.11 : l'élasticité selon type d'affaire (NB : affaire nouvelle et RNW : renouvellement)..... 27	
Figure 2.1 : l'évolution du taux de rétention par exercice..... 36	
Figure 2.2 : Répartition de la population selon le sexe	37
Figure 2.3 : Taux de rétention selon le sexe..... 37	
Figure 2.4 : Répartition des contrats conclus selon leur type d'échéance	38
Figure 2.5 : Taux de rétention selon type d'échéance..... 38	
Figure 2.6 : Répartition des primes selon le type de garantie..... 39	
Figure 2.7 : Taux de rétention selon type de garantie	39
Figure 2.8 : Taux de rétention en fonction de la prime	40
Figure 2.9 : Taux de rétention en fonction de l'ancienneté du contrat	41
Figure 2.10 : Taux de rétention en fonction de l'âge du conducteur	41
Figure 2.11 : Taux de rétention en fonction de CRM	42
Figure 3.1: Exemple de courbe de ROC	51
Figure 3.2 : Distribution de la probabilité de renouvellement estimée	59
Figure 3.3: Courbes ROC pour les bases apprentissage et test..... 61	
Figure 3.4 : Elasticité-prix en fonction de la probabilité de renouvellement	66
Figure 3.5 : Probabilité de rétention selon différents profils de risque	67
Figure 3.6 : Elasticité-prix selon différents profils de risque	69
Figure 3.7 : Aperçu sur l'optimisation	70

Chapitre I : Contexte de l'étude

I. Présentation d'organisme d'accueil :

1. AXA au monde :

Le groupe AXA est l'un des premiers groupes mondiaux d'assurance et de gestion d'actifs. Présent dans 64 pays, 165 000 collaborateurs d'AXA s'engagent aux côtés de 107 millions de clients. C'est la première marque mondiale d'assurance en 2017 selon le classement Best Global Brands établi par Interbrand.

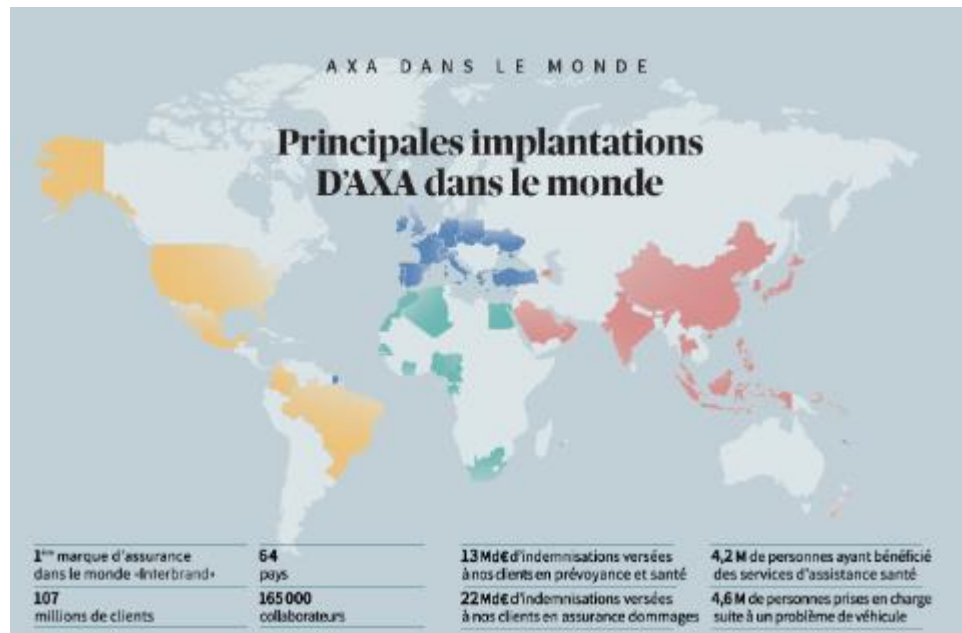


Figure 1.1 : Répartition géographique du groupe AXA dans le monde

Les activités d'AXA sont géographiquement diversifiées, avec une concentration sur les marchés d'Europe, d'Amérique du Nord et de la région Asie-Pacifique.

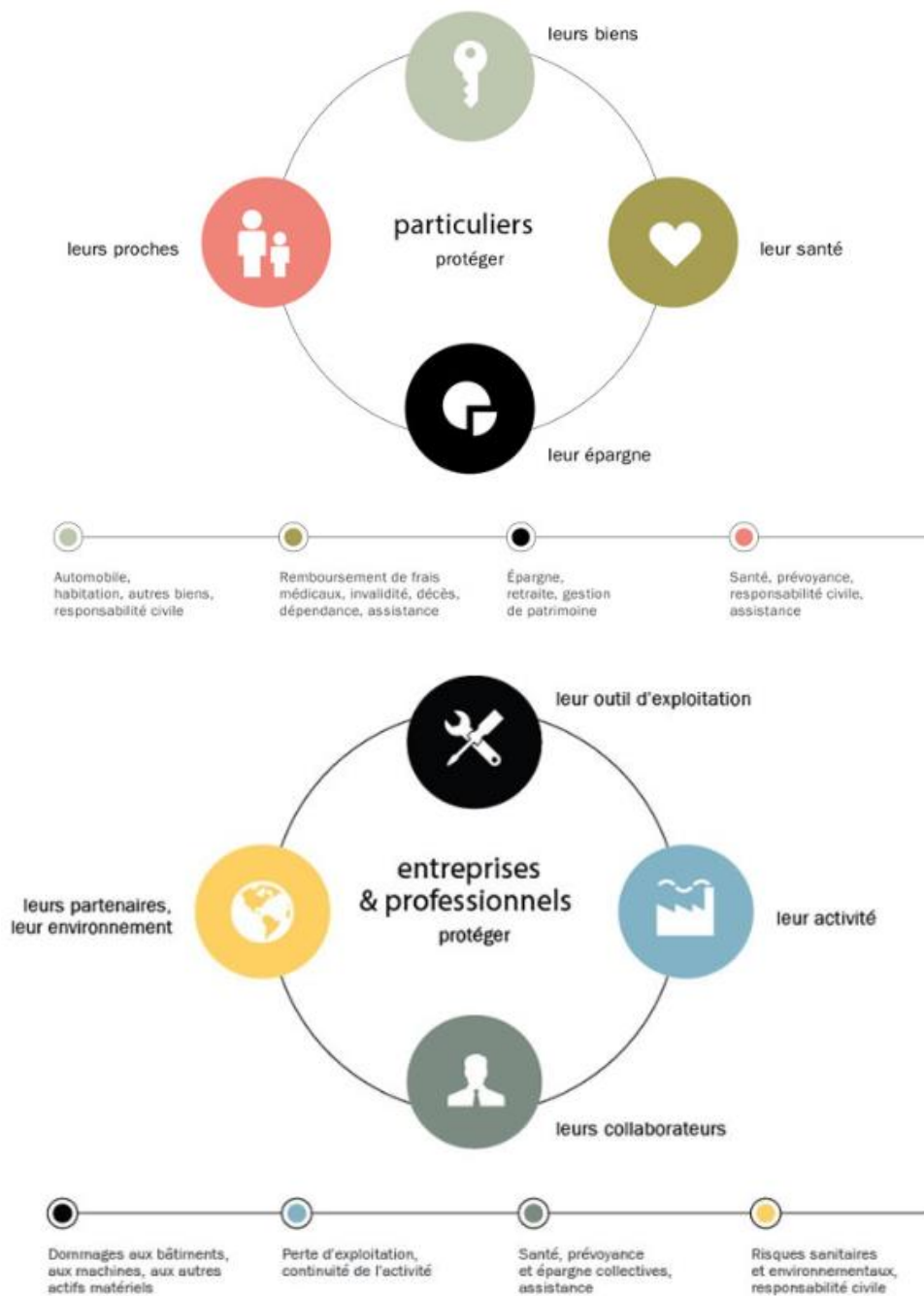
Le cœur de métier d'AXA est de proposer différentes solutions d'assurance à ses clients (particuliers, professionnels, entreprises ou institutions). Le groupe est spécialisé dans plusieurs domaines d'activité: assurance dommages, assurance de personnes (santé, prévoyance, épargne et retraite), gestion d'actifs, assistance, banque et protection juridique.

En 2019, AXA a réalisé un chiffre d'affaires de 103.5 milliards d'euros, soit une croissance de 5% en un an, le résultat net de cette année a atteint 3.9 milliards d'euros. Le ratio de solvabilité II s'établit à 198%, en hausse de 5 points par rapport au 31 décembre 2018.

2. AXA Assurance Maroc :

a. Présentation :

AXA Assurance Maroc (AAM) est la filiale marocaine du groupe AXA. Elle place le client au cœur de ses actions en aidant les particuliers à vivre plus confiant chaque étape de leur vie, et les entreprises et les professionnels à entreprendre plus sereinement.



L'ambition d'AAM est de devenir la société préférée dans son secteur d'activité, elle a remporté le label « Elu Service Client » de l'année 2020.

Les valeurs du groupe AXA :



b. Organigramme :

Au 31 décembre 2016, l'organigramme d'AXA Assurance Maroc se présente comme suit :

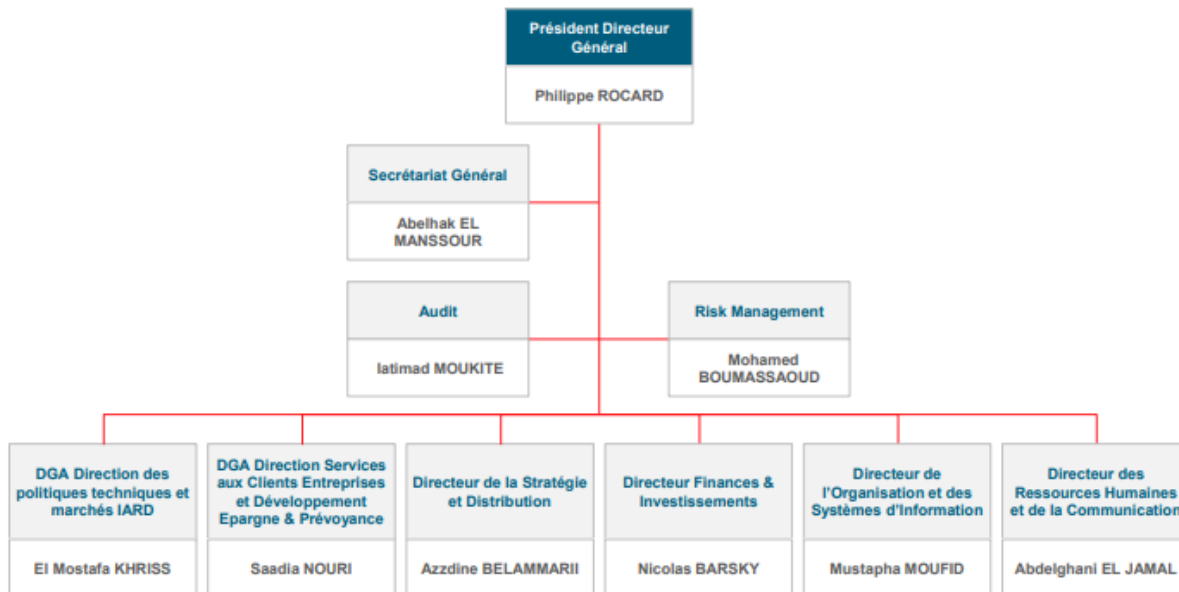


Figure 1.2 : Organigramme d'AAM

c. Produits commercialisés :

Les produits commercialisés pour chaque segment de clients sont énumérés ci-après (Consultez le site de la compagnie pour plus de détails) :

Segment	Offre	Produits
Particuliers	Epargne	Futuris Individuel II
		Educatis II
		Assurance Mixte
	Prévoyance	Essentiel Vie
		Familis
		Temporaire Décès
		Assurance Emprunteur
		Plan santé International - Individuel
	Auto	Solution Auto
		Solution Auto WW
		Solution Auto Fonctionnaire
	Habitation	Manzillouna
		Manzili
		Multirisque immeuble
	Moto	Scoters
		Cyclomoteurs
Moto		
Loisirs	Plaisance/Jet ski	
	RC chasse	
	Global Assistance Voyage	
Professionnels	Protection de l'activité	Multi Pro
		Multi PME-PMI
		Flotte Auto
		Multi Pharma
		Multi Hébergement
	TRC Awrach	
	Protection des collaborateurs	Prévoyance Santé Entreprise :
		Futuris Entreprise
		Accidents du Travail
		Plan Santé International
Entreprises	BTP	Tous Risques Chantier
	Industrie	Assurance Multirisque Industrielle
	Commerce et distribution	Contrat multirisque (Multi Industrielle et Marchandises Transportées) PME-PMI, Assurance Multirisque
	Services	Garanties de responsabilité civile

Tableaux 1.1 : Les produits d'assurance commercialisés par AAM

d. Historique :

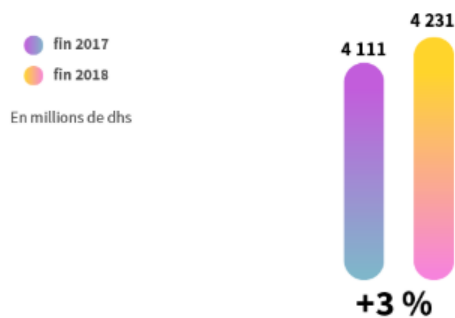
Le tableau ci-après est un récapitulatif de quelques évènements historiques qui ont marqué la création d'AMM :

1975	Création de la Société Al Amane
1993	Fusion de l'Entente (Non-vie) et Al Amane (Vie et santé)
1996	AXA fait l'acquisition d'UAP (Union des Assurances de Paris) et lance ses opérations au Maroc
1999	Al Amane devient AXA Al Amane
2000	AXA Al Amane devient AXA Assurance Maroc
2007	AXA devient actionnaire unique d'AXA Assurance Maroc

Tableau 1.2 : Historique d'AAM

e. Chiffres clés :

Chiffre d'affaires



Résultat net



Répartition du CA d'AAM à fin 2018

(en millions de dhs)

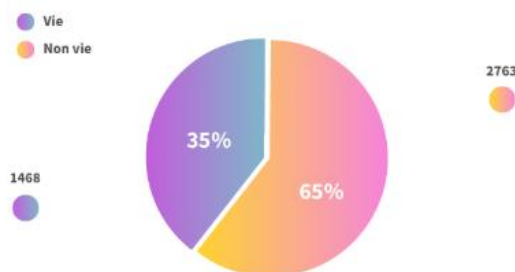


Figure 1.3 : Quelques chiffres clés sur la croissance d'AAM

II. Assurance automobile au Maroc :

1. Introduction :

L'assurance automobile comporte deux types de garanties :

Responsabilité Civile (RC) : elle est obligatoire et permet de couvrir la responsabilité civile du souscripteur du contrat. Ils sont couverts par cette garantie les dommages matériels et les dommages corporels. Depuis 2006 les compagnies ont acquis le droit de calculer eux-mêmes la prime sans intervention réglementaire, malgré la libéralisation des prix, le tarif RC est fixé à l'ancien niveau réglementaire.

Garanties Annexes : elles sont facultatives. En complément de la RC, les compagnies d'assurance proposent une panoplie de garanties permettant une meilleure protection (garantie incendie, vol, dommage tous accidents, dommage collision...), ces garanties annexes représentent 16% du CA automobile en 2018.

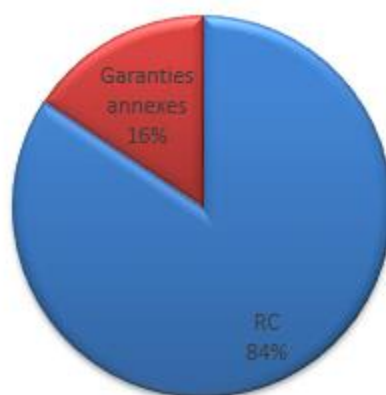


Figure 1.4 : Répartition des primes émises en assurance automobile selon le type de garantie en 2018

Les produits de la branche Automobile sont divers. Leur multitude émane de la diversité des risques couverts, de la couverture maximale que chaque compagnie détermine pour chaque produit et également des franchises que les entreprises d'assurance adoptent dans leurs normes de tarification. Nous résumons dans le tableau suivant les différentes garanties auto offertes par AXA Assurance Maroc en termes de risque couvert, de couverture maximale et de franchises adoptées par la compagnie :

Risque par sinistre	Couverture maximale	Franchises
Votre responsabilité et votre défense		
Responsabilité civile		
- Dommages corporels	50 000 000 Dh	
- Dommages matériels	50 000 000 Dh	
Défense et recours	Capital choisi figurant aux Conditions particulières	
Les dommages causés au véhicule		
Bris de Glaces (Article 6)	Valeur de remplacement ⁽¹⁾ dans la limite du capital assuré choisi	Taux figurant aux conditions particulières
<small>(1) Y compris frais de dépose et pose</small>		
Incendie (Article 7)		
- Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale	
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	
Vol (Article 8)		
- Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale	Taux figurant aux conditions particulières
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	
Evénements climatiques et naturels (Article 9)		
- Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale	Taux figurant aux conditions particulières
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	
Domage collision plafonnée (Article 10)		
- Véhicule et accessoires livrés par le constructeur	Capital choisi déclaré	Taux figurant aux conditions particulières
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	

Risque par sinistre	Couverture maximale	Franchises
Les dommages causés au véhicule		
Dommege collision déplafonnée (Article 10) - Véhicule et accessoires livrés par le constructeur _____ Valeur d'achat ou valeur vénale - Auto radio & remorque _____ Capital assuré déclaré - Aménagements professionnels _____ Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Dommege tous accidents (Article 11) - Véhicule et accessoires livrés par le constructeur _____ Valeur d'achat ou valeur vénale - Auto radio & remorque _____ Capital assuré déclaré - Aménagements professionnels _____ Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Dommege tous accidents éco plus (Article 12) - Véhicule et accessoires livrés par le constructeur _____ Capital assuré déclaré - Auto radio & remorque _____ Capital assuré déclaré - Aménagements professionnels _____ Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Dommege tous risques (Article 13) - Véhicule et accessoires livrés par le constructeur _____ Valeur d'achat ou valeur vénale - Auto radio & remorque _____ Capital assuré déclaré - Aménagements professionnels _____ Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Perte totale (Article 14) - Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale calculée selon le barème figurant aux conditions particulières	
Rachat de vétusté (Article 15)		
Les dommages causés aux personnes		
Protection familiale, conducteur et passagers (Article 20) - Capital Décès _____ } - Capital invalidité permanente _____ } - Frais médicaux _____ } Capitaux choisis figurant aux Conditions particulières		
Individuelle accidents conducteur habituel (Article 21) - Capital Décès _____ } - Capital invalidité permanente _____ } - Frais médicaux _____ } - Indemnité journalière d'hospitalisation _____ } Capitaux choisis figurant aux Conditions particulières		

Tableau 1.3 : les garanties auto offertes par AAM.

2. Tarif du produit automobile :

Le produit automobile d'AXA se présente sous forme de plusieurs formules : formule Tiers, formule Tiers Etendue, formule Accident et la formule Tous Risques. Le tableau ci-dessous résume les garanties offertes selon la nature de la formule :

	Tiers	Tiers +	Accident	Tous Risques
Responsabilité Civile	✓	✓	✓	✓
Protection familiale, conducteur et passagers	✓	✓	✓	✓
Incendie		✓	✓	✓
Vol		✓	✓	✓
Bris de glaces		En option	✓	✓
Toit panoramique				En option
Dommages Collision			✓	
Dommages tous accidents				✓
Rachat de vétusté	En option	En option	En option	En option
Individuelle Accidents du Conducteur	En option	En option	En option	En option
Evénements Climatiques et Naturels		En option	En option	En option
Perte Financière		En option	En option	En option
Inondation				En option
Valeur Majorée				En option
Aménagements professionnels		En option	En option	En option
Accessoires extérieurs				En option

Tableau 1.4 : Formules auto offertes par AAM.

La souscription à l'une des formules donne lieu au versement de la prime commerciale correspondante dont le montant varie en fonction des caractéristiques du conducteur et de son véhicule.

Le montant de la prime pure est le plus important dans la cotisation payée par le client. Cette partie est de plus complètement définie et incompressible. Les autres éléments de la prime commerciale sont décrits dans le graphique ci-contre :

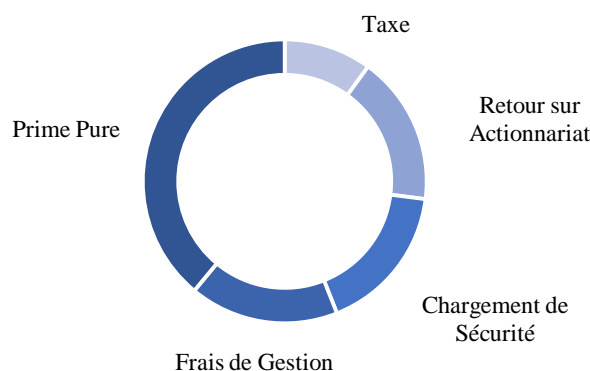


Figure 1.5 : Décomposition de la prime commerciale

Il s'agit :

- Des frais de gestion qui couvrent les frais de fonctionnement de la société,
- Des chargements de sécurité qui permettent à l'assureur de résister à la volatilité du business.
- Du retour auprès de l'actionnaire qui couvre la rémunération attendue par ces derniers.
- Des Taxes de la convention d'assurance.

3. Cycle de vie d'une police :

La vie d'une police est marquée principalement par trois événements :

- Émission d'une affaire nouvelle : l'assuré souscrit pour la première fois un contrat d'assurance chez une certaine compagnie ;
- Renouvellement : l'assuré a l'intention de renouveler son contrat pour la période suivante.
- Départ : l'assuré ne veut pas renouveler son contrat d'assurance parce qu'il veut changer de l'assuré ou il n'a plus besoin de la couverture. Le départ peut être dû à l'assuré (Résiliation ou Expiration du contrat) comme à l'assureur (Résiliation).

4. Vision globale du marché :

Le secteur d'assurance au Maroc est en plein essor, il présente un potentiel de développement important en termes d'offre et de volume. La branche automobile est une composante importante de ce secteur, elle s'accapare 27% du marché d'assurance et 48% de la partie non-vie, son chiffre d'affaires s'élève à 11 147,2 millions de dirhams en 2018 avec une croissance de 5.9% par rapport à l'année 2017. AXA en tant qu'acteur important dans le marché des assurances au Maroc (classé 4^{ème} en terme de CA), la partie automobile constitue 31.4% de son chiffre d'affaire en 2018.

L'assurance automobile au Maroc est un marché concurrentiel avec un certain nombre de compagnies très actives. Cette concurrence suscite les acteurs à lancer des produits de plus en plus attrayants avec des prix bien repensés. Le diagramme suivant présente la part de chaque compagnie de ce marché :

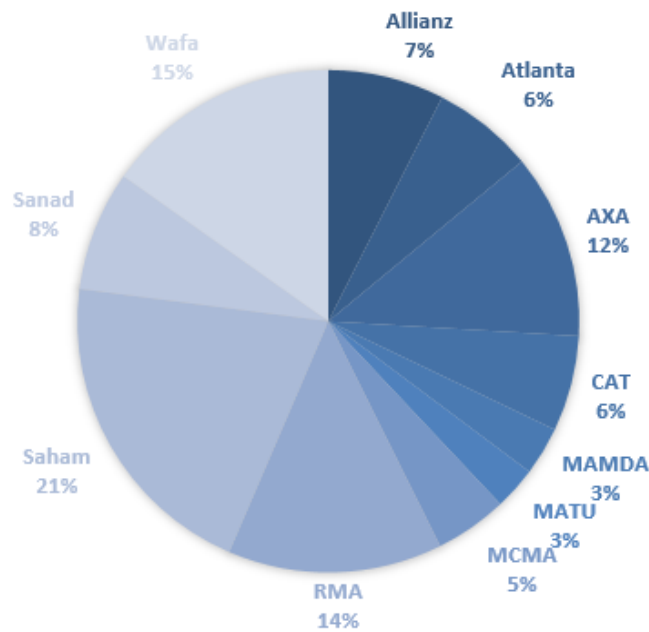


Figure 1.6 : Répartition des primes émises en automobile par entreprise d'assurance en 2018

III. Aperçu sur la tarification commerciale (commercial pricing) :

1. Contexte :

La tarification constitue l'un des cœurs de métier de l'actuariat. C'est une étape qui permet aux compagnies d'assurances d'évaluer les risques auxquels elles doivent faire face.

La réalisation d'un tarif (technique) en assurance IARD (auto, MRH, construction, etc.) s'appuie classiquement sur l'analyse de la charge de sinistre dans le cadre d'un modèle fréquence x coût dans lequel l'effet des variables explicatives sur le niveau du risque est modélisé par des modèles de régression de type GLM. Cette évaluation de risque néglige un élément essentiel, il s'agit du marché et sa dynamique (concurrence, comportement de l'assuré...). L'assureur doit procéder à une optimisation de ses tarifs proposés sur le marché en tenant en compte les primes offertes par ses concurrents afin d'élargir sa position dans le marché, maximiser sa marge et accélérer sa croissance.

2. Eléments de la tarification commerciale :

La figure ci-dessous montre les étapes à suivre pour réussir la tarification commerciale :

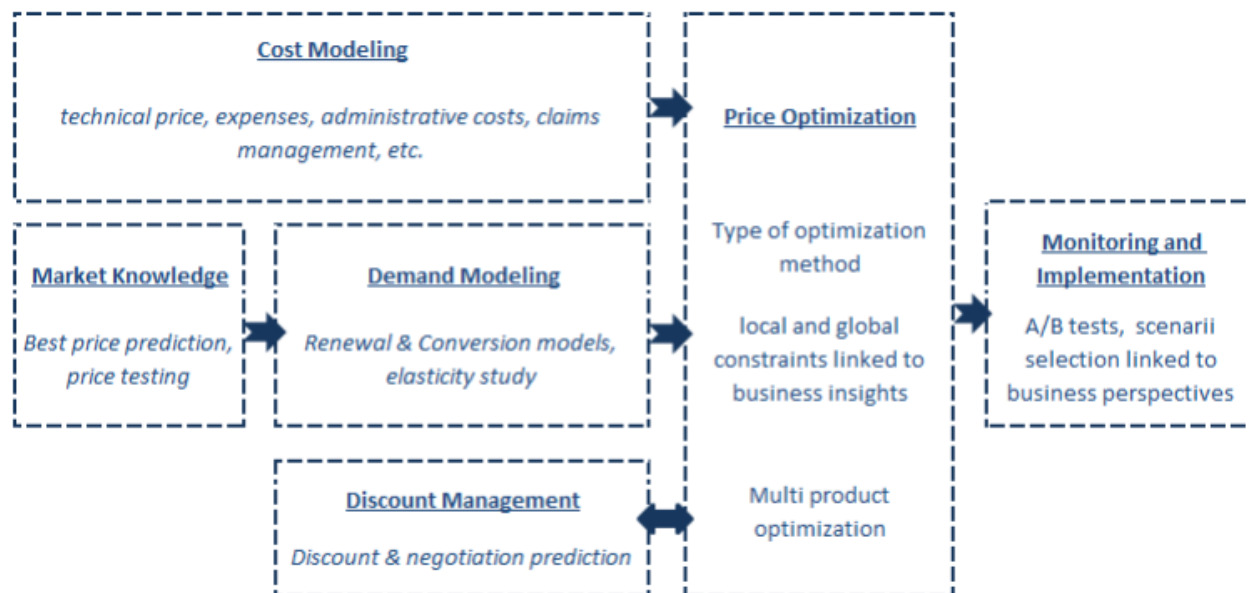


Figure 1.7 : Eléments de la tarification commerciale

On présentera dans la suite de ce chapitre les éléments clés de cette démarche de tarification, ils seront abordés en détail dans les prochains chapitres.

a. Connaissance du marché :

Une tarification technique (Cost Modeling) sophistiquée est une étape essentielle dans ce processus d'optimisation des tarifs, elle quantifie le risque que présente chaque contrat et permet d'échapper à l'antisélection, mais elle reste une approche insuffisante lorsque le marché connaît une forte concurrence. Le tarif à appliquer par l'assureur doit prendre en considération l'offre des concurrents et d'autres facteurs et mécanismes du marché.

i. Meilleur tarif (Best price) :

C'est le meilleur prix offert sur le marché pour un produit donné. La compétitivité influencera sans doute le choix de l'assuré : on constate d'après la figure suivante, élaborée par AXA Direct Poland, une baisse de la demande tant que le prix appliqué s'éloigne du meilleur tarif sur le marché (distance to Best Price : $dist_BP > 0$). Quand la prime demandée par cette entité est moins chère comparée à la meilleure prime dans le marché ($dist_BP < 1$), ceci attire un volume important de nouveaux de clients.

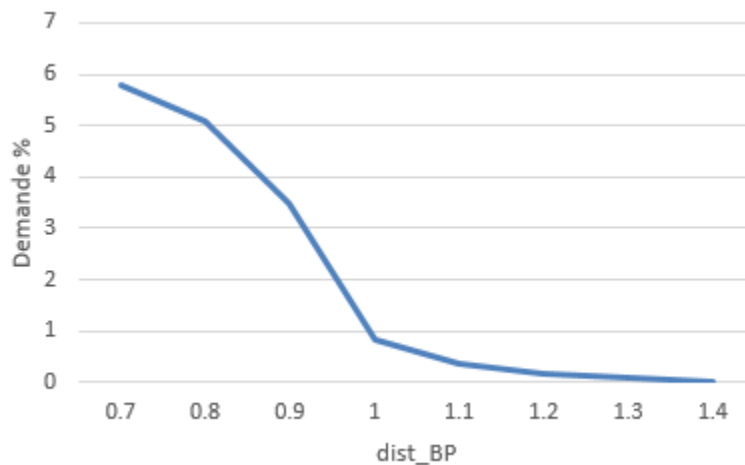


Figure 1.8 : Demande des affaires nouvelles en fonction de la distance par rapport au meilleur tarif

ii. Price testing :

C'est une expérimentation qui permettra à l'assureur de mieux comprendre le marché dont il fait partie, elle consiste à proposer des tarifs un peu différents pour des clients similaires (ayants le même profil de risque) afin d'observer leur réaction vis-à-vis de changement de prix et prédire leur comportement.

Le price testing est une approche empirique efficace pour tester la sensibilité du client à un changement de tarif. Il sera utile pour comparer ses résultats avec les prédictions du modèle de demande élaboré.



Figure 1.9 : Exemple de distributions utilisées dans le price testing

Cette distribution est recommandée lors de l'implémentation du price testing afin de générer des données fiables sur la sensibilité des clients aux différents niveaux de changement des prix tout en minimisant les coûts de cette expérimentation. On applique une variation (en %) positive et une variation (en %) négative similaire (égaux en valeur absolue) au même volume de clients.

b. Modélisation de la demande : (Voir chapitre III pour plus de détails)

La modélisation de la demande est l'élément clé de ce processus d'optimisation des tarifs. Un bon modèle de demande (élasticité) nous permettra de bien :

- ✓ Décrire et comprendre le comportement de notre client.
- ✓ Prédire l'impact d'un changement de tarifs sur l'attractivité de la compagnie et sa croissance.
- ✓ Optimiser davantage les tarifs à offrir sur le marché.

Modéliser la demande consiste à estimer la probabilité qu'un assuré souscrit un contrat pour la première fois ou sa probabilité de renouveler sa police s'il est déjà un client. D'après ce qui précède, deux types de modèles sont envisageables :

Modèle de conversion : ce modèle prédit la probabilité de conversion qui concerne la souscription des affaires nouvelles, il mesure l'attractivité de l'assureur dans le marché et sa capacité de séduire de nouveaux clients. Plusieurs facteurs peuvent influencer le choix d'un assuré : les prix, réputation et marque, qualité de services...

Modèle de rétention : on estime la probabilité (de rétention) qu'un client renouvelle son contrat d'assurance. Chaque compagnie d'assurance cherche à fidéliser ses clients en minimisant le taux des résiliations et garder un niveau de rétention élevé. Dans le présent mémoire, on se limitera à ce modèle. Une régression logistique sera utilisée, elle nous

permettra à approcher l'élasticité (sensibilité) de chaque client suite à un changement de tarif et déterminer les paramètres qui influencent la décision de l'assuré à renouveler / résilier son contrat.

Notion de l'élasticité-prix :

Élasticité de la demande mesure comment la demande d'un bien réagit après une variation de son prix.

$$Elasticité = - \frac{\frac{dD}{D}}{\frac{dP}{P}}$$

Le signe (-) est utilisé pour obtenir des valeurs positives de l'élasticité-prix puisqu'une augmentation de prix entrainera une baisse de la demande et vice versa (biens normaux).

$$Elasticité\ empérique = - \frac{\% \Delta D}{\% \Delta P} = - \frac{\frac{D_1 - D_0}{D_0}}{\frac{P_1 - P_0}{P_0}}$$

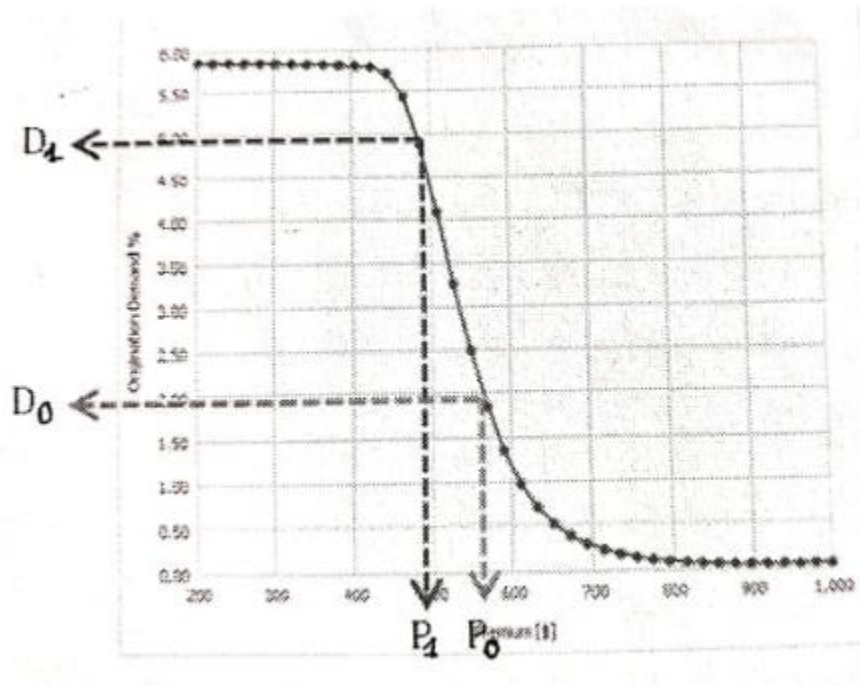


Figure 1.10 : Calcul de l'élasticité-prix empirique

Remarques :

- L'élasticité-prix n'est pas constante, elle change d'un assuré à un autre selon le profil de risque de chacun. Pour chaque segment de clients, on a une courbe de demande à partir de laquelle on peut calculer son élasticité par rapport à chaque niveau de prime.

- Généralement, les affaires nouvelles sont plus sensibles à une variation de prix que les renouvellements comme l'illustre la figure ci-dessous :

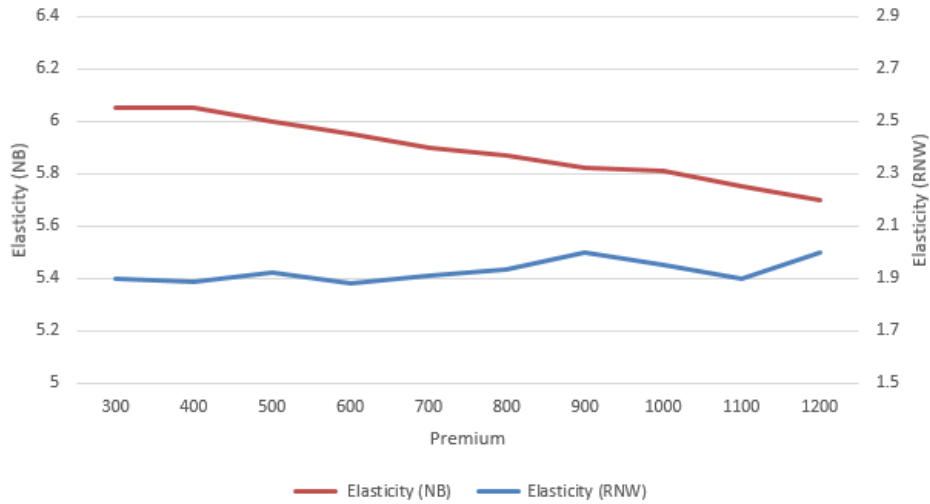


Figure 1.11 : l'élasticité selon type d'affaire (NB : affaire nouvelle et RNW : renouvellement)

c. Optimisation des primes (Price optimization) :

D'après l'enchaînement de cette stratégie de tarification décrite dans la figure 1.7, l'étape de l'optimisation demande un modèle technique sophistiqué et un bon modèle d'élasticité. L'optimisation des prix consiste à chercher la prime optimale à appliquer qui maximise la marge du portefeuille et sa croissance tout en respectant quelques contraintes budgétaires et sans impacter les orientations et les valeurs de l'assureur. Ci-après un exemple de programmes d'optimisation :

$$(P): \begin{cases} \max_{P_1, P_2, \dots, P_n} \sum_{i=1}^n (P_i - AC_i) * f(P_i, \mathbf{X}_i) \\ P_i \leq \overline{P}_i \quad \forall i \in \{1, 2, \dots, n\} & (1) \\ P_i \geq \underline{P}_i \quad \forall i \in \{1, 2, \dots, n\} & (2) \\ \frac{1}{n} \sum_{i=1}^n f(P_i, \mathbf{X}_i) \geq \gamma & (3) \end{cases}$$

Avec:

- P_i : Prime optimale associée au client i ;
- AC_i : Prime technique associée au client i ;
- $f(P_i, \mathbf{X}_i)$: Probabilité de renouvellement du client i estimée à partir du modèle de rétention ;

- \underline{P}_i et \overline{P}_i deux niveaux respectivement inférieur et supérieur de la prime fixée pour le client i ;
- γ : niveau minimum de rétention ciblé.

IV. Problématique :

L'assurance automobile au Maroc est actuellement confrontée à une concurrence très forte. Cette concurrence est principalement due à la multiplication des acteurs sur ce secteur. Dans ce contexte ultra-concurrentiel, les techniques de tarification traditionnelles basées principalement sur l'estimation du risque ne suffisent plus à relever les défis de croissance et de rentabilité demandée par l'actionnaire. Le modèle traditionnel optimise la prise de risque de l'assureur, mais pas son profit.

En réaction à cet environnement compétitif, les assureurs ont développé de nouvelles stratégies-prix basées aussi bien sur l'estimation de la prime pure (modèle technique traditionnel) que sur l'élasticité-prix des souscripteurs (modèle de la demande).

L'objectif de ce mémoire est d'étudier la rétention des clients en assurance automobile en modélisant leur probabilité de renouvellement. Cette dernière sera notre point de départ pour évaluer l'élasticité des assurés au prix et bien estimer leur comportement.

Chapitre II : Analyse exploratoire des données

Avant chaque démarche de modélisation, une étude d'exploration de la base de données s'avère nécessaire afin de la rendre plus exploitable et adéquate. Nous présenterons dans ce chapitre les différentes étapes suivies dans la construction et la préparation de notre base de données et une analyse exploratoire examinant les liaisons entre les variables jugées explicatives de notre variable cible.

I. Présentation des données manipulées :

Dans le processus de préparation de notre base de données à la modélisation, trois sources volumineuses de données ont été mises à notre disposition :

1. Base « Contrat » :

Cette base de données regroupe l'ensemble des polices en assurance automobile qui concernent les véhicules à usage tourisme souscrites chez AXA Assurance Maroc et ayant effet entre la période 2013-2019. Soit 4.5 millions lignes dont chacune représente un avenant spécifique sur un contrat donné. Le tableau suivant regroupe l'ensemble des variables formant cette base :

Nom de la variable	Désignation
Polnum	Code police
Polavn	Code avenant
Typanv	Type avenant
Stpavn	Sous type avenant
Codtef	Date d'effet
Codtex	Date d'expiration (échéance)
Codur	Code durée (l'exposition du contrat en mois)
Cotyec	Type d'échéance
Codefc	Date premier effet du contrat

Tableau 2.1 : Description de la base « Contrat »

2. Base « Conducteur » :

Ce tableau de données nous informe sur le conducteur principal dans chaque contrat. Il est structuré de la même façon que la base « Contrat », chaque ligne enregistre les modifications dans les caractéristiques du conducteur suite à l'émission d'un certain avenant. Les colonnes structurant la base « Conducteur » sont comme suit :

Nom de la variable	Désignation
Polnum	Code police
Polavn	Code avenant
Dunom	Nom
Duprno	Prénom
Ducin	CIN
Dusex	Sexe
Dudtna	Date de naissance
Dudtpr	Date obtention du permis
Vicvil	Code ville
Dusitf	Situation familiale (état matrimoniale)
Pfcprf	Code profession
Dutyp	Type conducteur

Tableau 2.2 : Description de la base « Conducteur »

3. Base « Véhicule » :

Cette base de données ressemble à la base « conducteur », elle nous renseigne sur les véhicules enregistrés sur chaque contrat conclu. Voici les variables composant cette base :

Nom de la variable	Désignation
Polnum	Code police
Polavn	Code avenant
Venmob	Matricule du véhicule
Veenr	Energie (Essence/Gasoil)
Vepui	Puissance fiscale
Vecyli	Cylindrée
Vedcir	Date de mise en circulation
Vetxcr	Coefficient Réduction/Majoration
Veplc	Nombre de places

Vecarr	Carrosserie
Vealn	Valeur neuf
Vealv	Valeur vénale
Vemarq	Marque

Tableau 2.3 : Description de la base « Véhicule »

II. Préparation de la base de données :

1. Création de la variable cible :

On a utilisé dans cette étape la base « Contrat » qui nous renseigne sur chaque contrat souscrit : sa date d'effet, date d'expiration, type d'échéance (Durée Ferme (DF) ou Durée Compagnie (DC))...Chaque ligne de ce tableau correspond à un avenant, il s'agit d'une modification apportée au contrat suite à un changement de situation de l'assuré tel qu'un changement de véhicule, ajout de garanties...il existe 17 catégories d'avenants d'assurance automobile chez AXA.

Lors de la souscription d'un contrat d'assurance, l'agent demande à l'assuré de choisir entre un contrat de :

- Durée Ferme (DF): un contrat qui s'expire simplement à sa date d'échéance.
- Durée Compagnie (DC): le contrat se renouvelle automatiquement à chaque date d'anniversaire.

On rappelle que notre variable réponse (Y) est binomiale, elle ne peut prendre que deux valeurs :

$$Y = \begin{cases} 1 & \text{l'assuré renouvelle son contrat} \\ 0 & \text{sinon} \end{cases}$$

Notre premier souci était de créer avec soin cette variable puisque c'est celle qui va quantifier la probabilité de renouveler ou résilier un contrat d'assurance et l'élasticité-prix de chacun de nos clients.

Pour (Y=1) : nous avons sélectionné les cinq avenants de renouvellement à savoir :

- ✓ (typanv=2 & stpavn=1) : avenant de renouvellement DF à DF ;
- ✓ (typanv=2 & stpavn=2) : avenant de renouvellement DF à DC ;
- ✓ (typanv=2 & stpavn=3) : avenant de renouvellement DC à DF;
- ✓ (typanv=2 & stpavn=4) : avenant de renouvellement DC à DC ;
- ✓ (typanv=2 & stpavn=5) : réémission d'avenant de renouvellement.

Pour (Y=0) : nous avons considéré les résiliations (avenants typanv=6) émanant de l'assuré en vue de construire un modèle captant les décisions prises volontairement par notre client de renouveler ou résilier son contrat d'assurance.

Une fois nos avenants d'intérêt sont sélectionnés, nous avons procédé à une vérification de leur chronologie en faisant un regroupement par police (polnum) qui assure un ordre croissant des avenants selon leurs dates d'effet (codtef). Cela nous a permis de détecter plein de problèmes de cohérence à traiter, on cite ci-dessous des cas de figure :

- ✓ Dans un même contrat on trouve deux avenants de renouvellement ayant la même date d'effet. Si l'un des deux avenants est de type 2-5 (réémission d'avenant de renouvellement) on se débarrasse automatiquement de l'autre, sinon on compare la date d'effet de l'avenant suivant avec les dates d'expiration des deux avenants. Si de plus ces derniers sont égaux on peut se baser sur la variable type d'échéance en regardant l'avenant précédant pour sélectionner le correct avenant.

POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC
13	76 2014G318167	2	1	20140411	DF	06	20141010	20121011
13	76 2014G392573	2	5	20140411	DF	03	20140710	20121011
13	76 2014G542772	2	1	20140711	DF	06	20150110	20121011

⇒ Avenant 2-5 sera retenu. En regardant la date d'effet du prochain avenant de renouvellement, on peut s'assurer que nous avons sélectionné le bon avenant.

POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC
101	81 2014G514806	2	3	20140609	DF	02	20140808	20120609
101	81 2014T090651	2	4	20140609	DC	A0	20150608	20120609
101	81 2014F219176	6	8	20140808	DF	02	20140808	20120609

⇒ L'avenant 2-3 sera retenu.

POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC
11	31 2013C027222	2	1	20130103	DF	10	20140102	20121201
11	31 2014G111517	2	3	20140121	DF	10	20150120	20121201
11	31 2014G074560	2	2	20140121	DC	A0	20150120	20121201

⇒ L'avenant 2-2 sera retenu puisque l'avenant de renouvellement 2-3 suppose que le contrat avait une échéance de type DC.

- ✓ On a remarqué dans certaines polices que plusieurs avenants d'expiration et de résiliation ont été sélectionnés, parce que pour les contrats de type DF l'avenant d'expiration (6-8) s'ajoute automatiquement quand le contrat arrive à sa date d'échéance et on peut aussi envisager le cas d'un départ d'un assuré puis il revient renouveler son contrat après un certain temps. Dans telles situations, on a retenu le dernier avenant de résiliation ou d'expiration s'il n'est pas suivi par un avenant de renouvellement.

POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC
13	07 2013G085219	2	1	20131003	DF	03	20140102	20121003
13	07 20140004917	6	8	20140102	DF	03	20140102	20121003
13	07 2014G051936	2	1	20140111	DF	03	20140410	20121003
13	07 2014F105385	6	8	20140410	DF	03	20140410	20121003
13	07 2014G321350	2	1	20140412	DF	03	20140711	20121003
13	07 2014F181087	6	8	20140711	DF	03	20140711	20121003

⇒ Le dernier avenant d'expiration (6-8) sera retenu.

POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC
359	61 2018C197185	1	1	20180322	DF	12	20190321	20180322
359	61 2019F119276	6	8	20190321	DF	12	20190321	20180322
359	61 2019C328162	2	1	20190401	DF	12	20200331	20180322

⇒ Dans ce cas l'avenant d'expiration (6-8) ne sera pas considéré.

- ✓ Il existe des assurés qui ont résilié leurs contrats et reviennent après une période donnée souscrire une nouvelle police sur le même véhicule sans mentionner à l'agent qu'ils étaient déjà parmi les clients de la compagnie (leur numéro de police change). Pour pallier à ce problème, nous avons utilisé le CIN du conducteur (Ducin) et la matricule du véhicule (Venmob) afin détecter les différents codes police (polnum) qui correspond au même contrat. Une fois résolu, un avenant de résiliation sera négligé et l'avenant relatif à l'émission d'une affaire nouvelle sera considéré comme renouvellement.

Les deux codes avenants suivants correspondent au même contrat, on sélectionne un seul polnum et on remplace l'autre.

POLNUM	VENMOB	DUCIN
39 23	1-A-01	E4 91
259 88	1-A-01	E4 91

POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC
259	88 2014G093591	1	1	20140128	DF	03	20140427	20140128
259	88 2014F121305	6	8	20140427	DF	03	20140427	20140128
39	23 2014C204274	1	1	20141106	DF	03	20150205	20141106
39	23 2015F045708	6	8	20150205	DF	03	20150205	20141106



POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC	y
39	23 2014C204274	1	1	20141106	DF	03	20150205	20141106	1
39	23 2015F045708	6	8	20150205	DF	03	20150205	20141106	0

⇒ L'avenant (6-8) ayant comme date d'effet le 27/04/2014 est négligé et l'avenant (1-1) qui correspond à l'émission d'une affaire nouvelle à la date 06/11/2014 a été considéré comme renouvellement.

2. Introduction des variables externes à notre base de données :

a. Prime :

La prime qu'a payé chaque assuré constituera une variable explicative clé dans notre modèle d'élasticité. Nous l'avons introduit à partir d'une autre base de données avec une autre variable « Equipement » indiquant la garantie achetée par l'assuré parmi la responsabilité civile (RC), dommage collision (DC) ou dommage tous accidents (DTA).

b. Ecart à la concurrence (Best price) :

Cette variable semble très intéressante à notre modèle de demande, c'est la meilleure prime que trouvera un client sur le marché. Elle sera tirée d'un benchmark de tarifs proposés par les cinq premiers concurrents d'AXA dans la branche automobile.

On définit l'écart à la concurrence « dist_BP » comme la différence entre la prime et le tarif le moins cher proposé sur le marché (Best Price) :

$$\begin{cases} \text{dist_BP}_i > 0 & \text{AXA demande une prime plus chère au segment } i \\ \text{dist_BP}_i \leq 0 & \text{Le tarif proposé par AXA au segment } i \text{ est le meilleur sur le marché} \end{cases}$$

Le benchmark utilisé n'est pas exhaustif, il ne couvre pas tous les segments possibles d'assurés puisque chaque assureur utilise son propre modèle de tarification et les segmentations du risque adoptées diffèrent.

c. Autres variables :

Afin d'enrichir notre modèle, nous avons cherché d'autres variables externes de dimension socio-économique susceptibles d'expliquer le comportement de notre client. On cite ci-après les variables collectées :

- ✓ Pourcentage des intermédiaires d'AXA dans chaque région ;
- ✓ Densité des véhicules en circulation dans chaque région ;
- ✓ Taux d'activité et de chômage dans chaque région et selon le sexe ;
- ✓ Nombre de véhicules par ménage dans chaque région ;
- ✓ Dépense annuelle moyenne par ménage.

Toutes les variables (à l'exception de la première) ont été tirées des annuaires statistiques officielles publiées sur le site du haut-commissariat au plan.

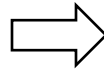
3. Valeurs manquantes et aberrantes :

Quelque avenants sélectionnés ne se trouvent pas dans l'une des deux bases « Conducteur » et « Véhicule », dans ce cas on a copié les informations manquantes de l'avenant ayant la date d'effet la plus proche.

POLNUM	POLAVN	TYPANV	STPAVN	CODTEF	COTYEC	CODUR	CODTEX	CODEFC	DUDTPR	DUDTNA	VICVIL	DUSITF	PFCPRF	DUSEX	DUCIN
13	26 2013G109109	2	1	20131010	DF	03	20140109	20121005							
13	26 20140014170	6	8	20140109	DF	03	20140109	20121005	20120419	19800101	615 C		99999 M	WA	53

Le tableau suivant montre un exemple de valeurs non fiable ou aberrantes que nous avons détecté dans cette phase d'épuration de la base de données. Ces valeurs ont été traitées comme des manquantes puis imputées par la moyenne (arrondie en un nombre entier) :

Variable	Minimum	Maximum
age_conducteur	-6180.00	375.0000000
age_permis	-3.0000000	119.0000000
age_vehicule	-2.0000000	99.0000000
ancienete_contrat	0	62.0000000



Variable	Minimum	Maximum
age_conducteur	18.0000000	90.0000000
age_permis	0	51.0000000
age_vehicule	0	47.0000000
ancienete_contrat	0	14.0000000

III. Analyse exploratoire des données :

Nous examinerons dans cette partie les différentes relations liant notre endogène avec les différentes variables explicatives, nous ferons aussi une diagnostique de colinéarité pour s'assurer de l'indépendance des variables introduites à notre modèle de rétention.

On utilisera dans la suite de notre exploration de données la définition suivante du taux de rétention : « Taux de rétention annuel est une moyenne pondérée de Y par l'exposition annuelle de chaque contrat »

La motivation derrière cette pondération s'illustre dans l'exemple suivant : Considérons un contrat qui a été renouvelé pour une durée d'une année (une seule ligne Y=1) est équivalent 4 renouvellements successives d'un autre contrat ayant une exposition trimestrielle (4 lignes de Y=1). Dans ce cas la moyenne arithmétique de Y n'est pas appropriée.

Le graphique suivant trace l'évolution de la rétention dans le temps. On note un taux stable de rétention d'un exercice à un autre, chaque année environs de 77% des assurés de la branche auto chez AXA Maroc restent fidèles à ce label, Or 23% des clients quittent annuellement la compagnie pour des multiples raisons parmi lesquelles on peut citer :

- Le client n'est plus satisfait du service (le délai et la façon de gérer le sinistre, remboursement, ...)
- Le tarif qu'on lui demande n'est plus le meilleur sur le marché ;
- D'autres facteurs externes liés à la situation socio-économique du client : la vente de véhicule, chômage ou changement d'emploi...

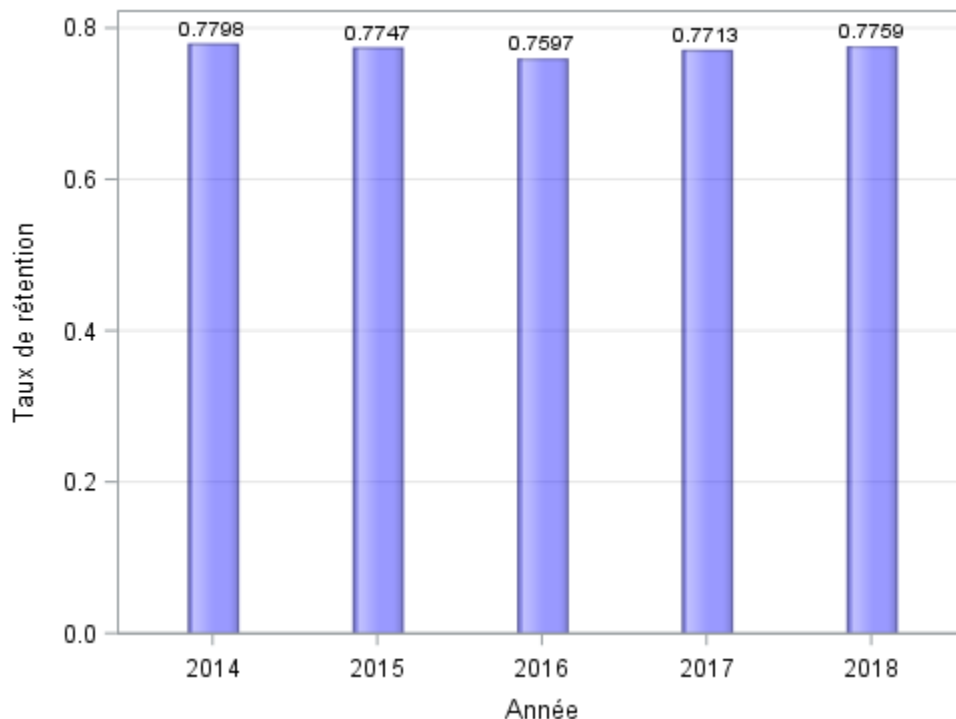


Figure 2.1 : L'évolution du taux de rétention par exercice

On peut envisager autres cas de figures où se manifeste la rétention. Il existe des clients qui mettent fin à leurs contrats, puis ils reviennent après un certain temps (2 ans, 3 ans,...) pour renouveler, dans ce cas la probabilité que telles personnes reviennent pour souscrire une nouvelle police reste faible par rapport à la rétention annuelle.

1. Taux de rétention selon le sexe :

88% des clients d'AXA ayant acheté un produit d'assurance auto dans la période 2013-19 sont des hommes :

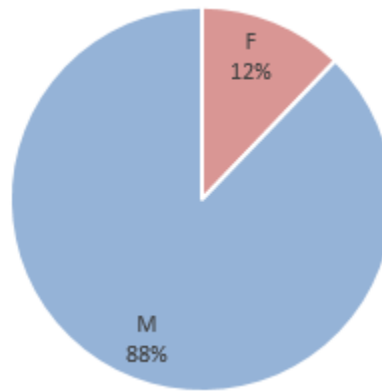
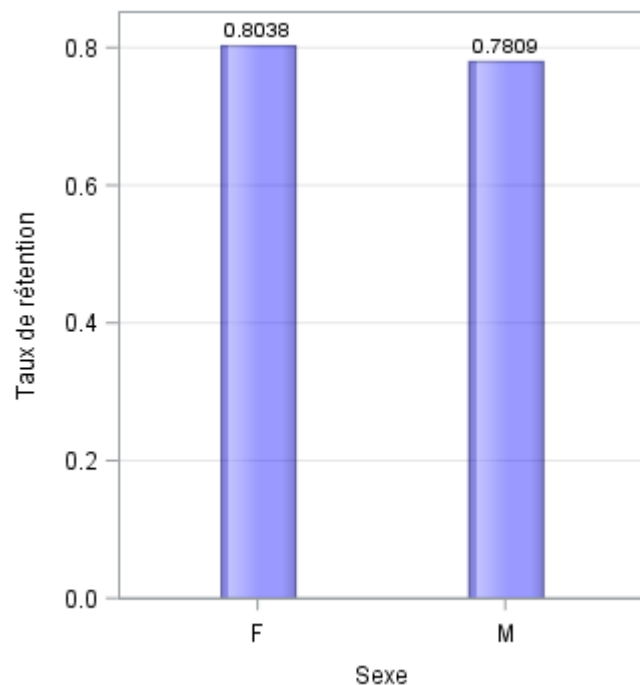


Figure 2.2 : Répartition de la population selon le sexe

On constate d'après la figure ci-après une différence entre les deux sexes dans leur choix de renouveler/résilier un contrat d'assurance automobile conclu chez AXA Assurance Maroc. Les femmes sont plus fidèles que les hommes : le test (asymptotique) d'indépendance de khi-deux affirme cette constatation à 5% près.



Statistics for Table of Y by DUSEX

Statistic	DF	Value	Prob
Chi-Square	1	4008.5139	<.0001
Likelihood Ratio Chi-Square	1	4158.7618	<.0001
Continuity Adj. Chi-Square	1	4008.2217	<.0001
Mantel-Haenszel Chi-Square	1	4008.5122	<.0001
Phi Coefficient		-0.0410	
Contingency Coefficient		0.0410	
Cramer's V		-0.0410	

Sample Size = 2737747

Figure 2.3 : Taux de rétention selon le sexe

2. Taux de rétention selon type d'échéance :

La majorité des contrats souscrits composant notre portefeuille d'étude ont une durée ferme :

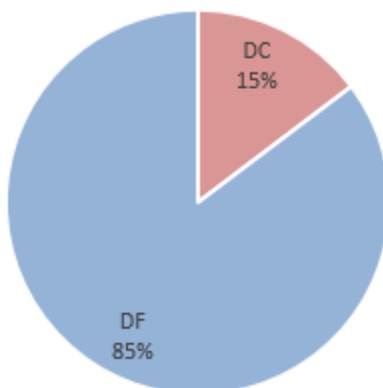
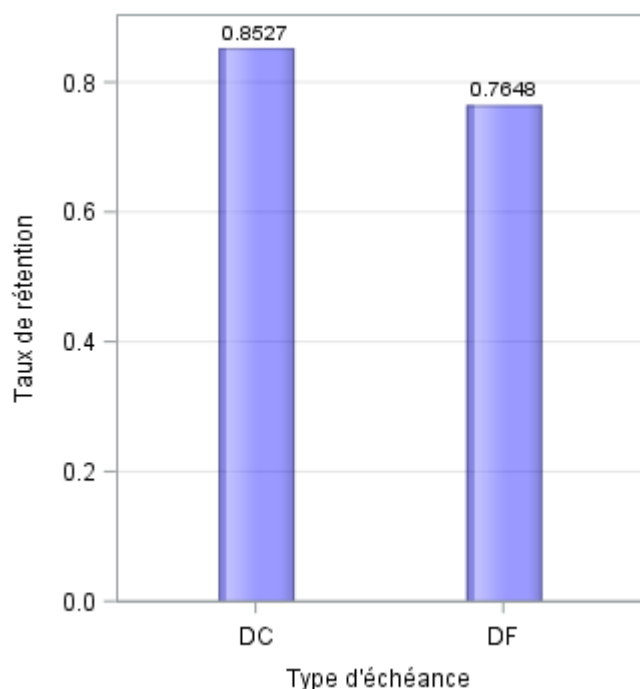


Figure 2.4 : Répartition des contrats conclus selon leur type d'échéance

On regarde sur la figure ci-après que les contrats dont l'échéance est de type DC ont un taux de rétention élevé par rapport aux contrats DF, ce qui était prévisible puisque les gens ayant souscrit des contrats DC avaient une intention préalable de renouveler leurs polices.



Statistics for Table of Y by COTYEC			
Statistic	DF	Value	Prob
Chi-Square	1	35513.9651	<.0001
Likelihood Ratio Chi-Square	1	39453.7498	<.0001
Continuity Adj. Chi-Square	1	35513.2005	<.0001
Mantel-Haenszel Chi-Square	1	35513.9502	<.0001
Phi Coefficient		-0.1220	
Contingency Coefficient		0.1211	
Cramer's V		-0.1220	

Figure 2.5 : Taux de rétention selon type d'échéance

3. Taux de rétention selon type de garantie :

Ci-dessous la distribution des différentes primes émises dans le cadre de notre portefeuille étudié selon la variable « Equipement » :

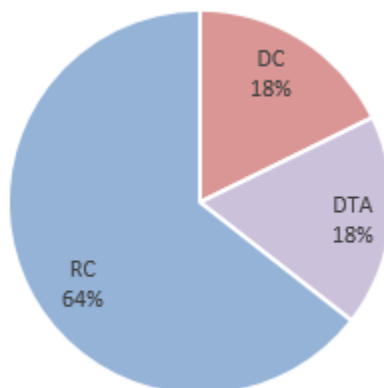
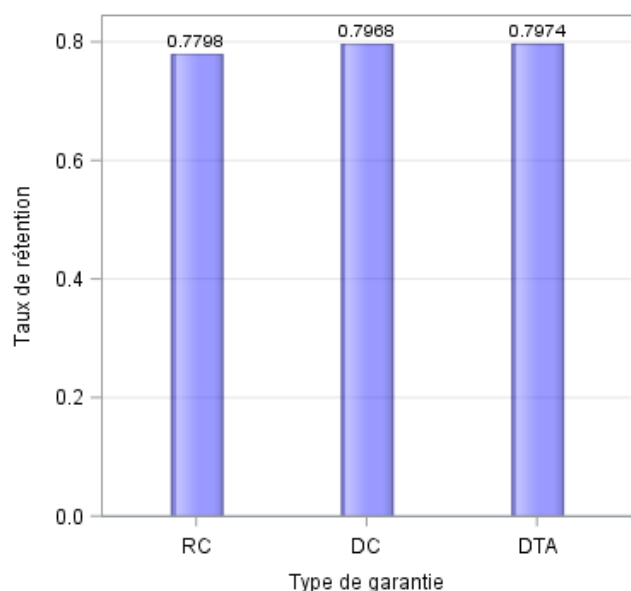


Figure 2.6 : Répartition des primes selon le type de garantie



Statistics for Table of Y by EQUIPEMENT			
Statistic	DF	Value	Prob
Chi-Square	2	13071.0497	<.0001
Likelihood Ratio Chi-Square	2	13749.7145	<.0001
Mantel-Haenszel Chi-Square	1	11723.3226	<.0001
Phi Coefficient		0.0740	
Contingency Coefficient		0.0738	
Cramer's V		0.0740	

Figure 2.7 : Taux de rétention selon type de garantie

Les gens ayant souscrit la garantie DTA ou DC sont plus probables à renouveler leurs contrats par rapport aux assurés qui détiennent seulement la responsabilité civile (RC).

4. Taux de rétention en fonction de la prime :

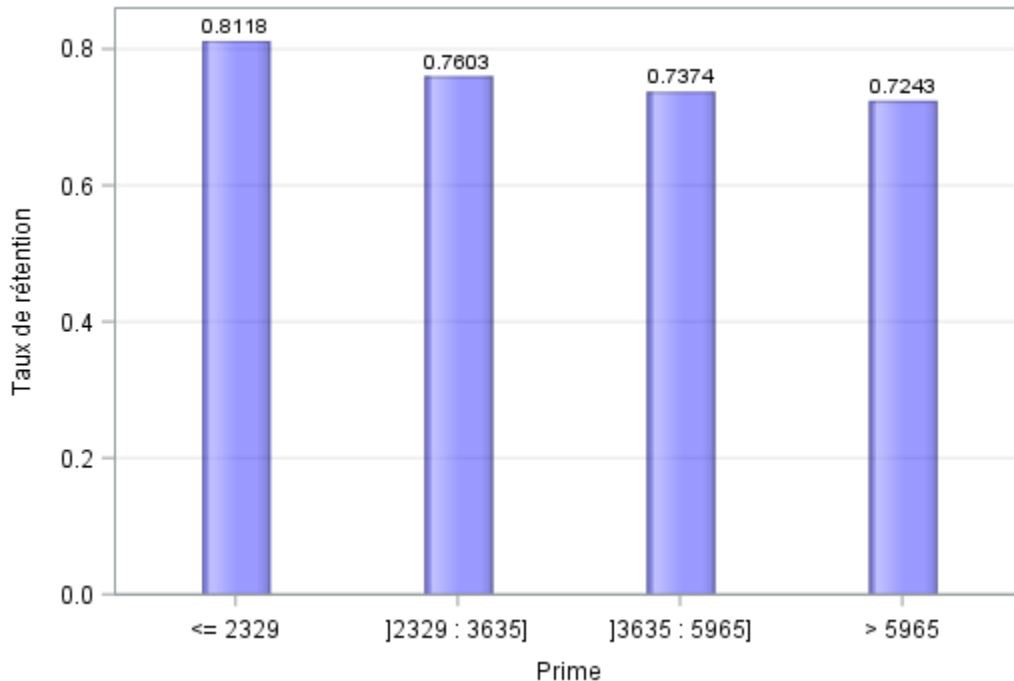


Figure 2.8 : Taux de rétention en fonction de la prime

On remarque clairement la baisse du taux de rétention avec l'augmentation de la prime, on retrouve un tel constat dans le cas des biens normaux sous l'hypothèse d'un consommateur rationnel : la demande du bien décroît suite à une hausse de prix. Quantifier l'effet de la prime sur la rétention nous donnera une idée sur l'élasticité-prix de nos clients, on prendra ainsi les bonnes décisions sur les tarifs à offrir en optimisant la rentabilité de notre portefeuille.

5. Taux de rétention en fonction de l'ancienneté du contrat :

La figure suivante montre que les clients les plus anciens ont moins de chance de résilier. Cela s'explique par le fait qu'un client fidèle ne s'intéressera pas aux prix des concurrents. Étant peu informés des tarifs du marché, ils ne réagissent pas fortement aux évolutions tarifaires.

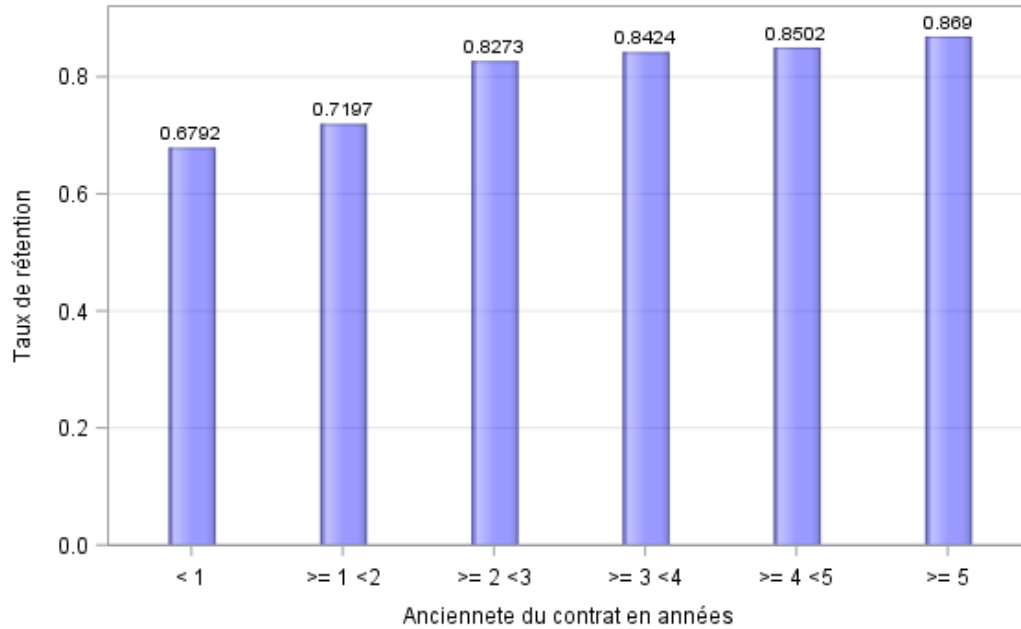


Figure 2.9 : Taux de rétention en fonction de l'ancienneté du contrat

6. Taux de rétention en fonction de l'âge du conducteur :

Les clients âgés de moins de 25 ans ont une probabilité de résiliation élevés par rapport aux autres tranches d'âge. Ce résultat est logique vu que les jeunes disposent généralement de revenus modestes et qu'ils sont très attentifs aux offres disponibles sur le marché.

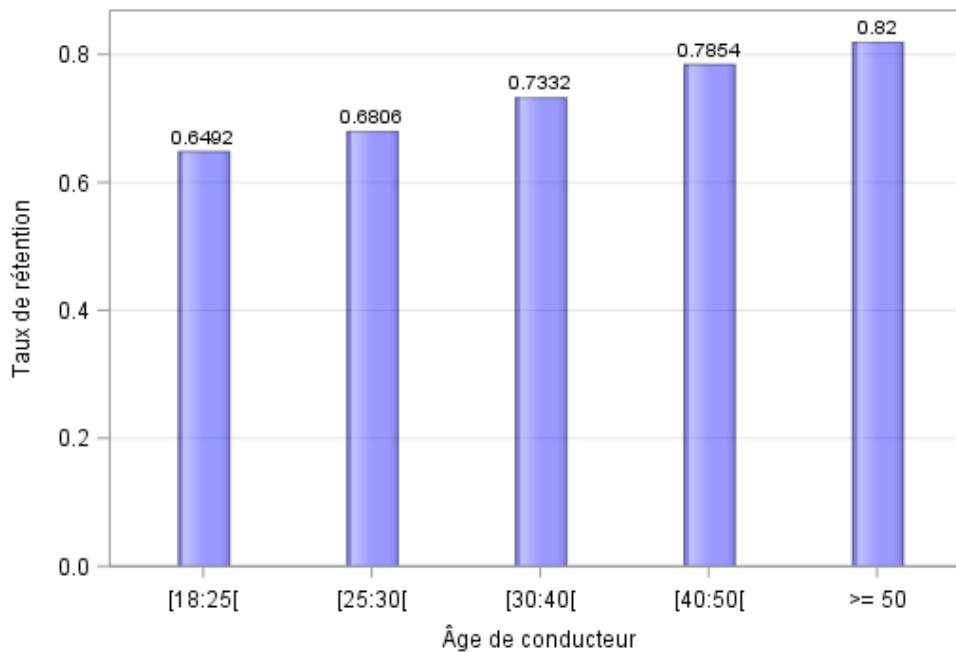


Figure 2.10 : Taux de rétention en fonction de l'âge du conducteur

7. La rétention en fonction du taux CRM :

Le Coefficient de Réduction / Majoration est un coefficient multiplicateur qui s'applique sur la prime d'assurance automobile compte tenu de l'historique de sinistralité. Il permet d'octroyer une réduction sur la prime d'assurance aux bons conducteurs et de majorer les primes des assurés ayant causé des sinistres. Plus précisément, il fonctionne comme suit :

- Réduction de 10% de la prime si l'assuré n'a causé aucun accident engageant totalement ou partiellement sa responsabilité durant une période d'assurance de 24 mois consécutifs précédant la souscription ou le renouvellement de son contrat ;
- En cas d'un sinistre engageant totalement ou partiellement la responsabilité du conducteur enregistré durant les 12 mois de la souscription, le renouvellement de l'assurance est majoré de 20% pour chaque accident matériel et 30% pour un sinistre corporel sans pour autant pouvoir dépasser 250% de la prime de base.

Le digramme ci-dessous montre la répartition de la rétention selon le Coefficient Réduction/Majoration :

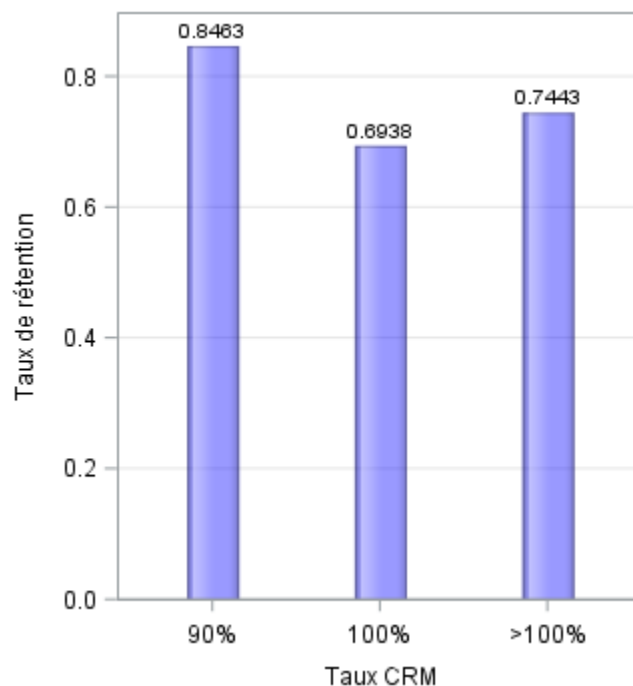


Figure 2.11 : Taux de rétention en fonction de CRM

Toute compagnie d'assurance cherche à fidéliser ses clients ayant un bon profil de risque et éviter l'antisélection, c'est ce qu'on constate sur le graphique ci-dessus. On observe un taux de rétention assez élevé chez les bons conducteurs (CRM=0.9).

La rétention chez les gens ayant un taux CRM de 100% est faible comparée aux deux autres groupes, ces nouveaux conducteurs sont généralement des jeunes, ils auront donc un taux de résiliation important.

Conscients d'être des mauvais conducteurs (CRM > 100%), ces clients sont plus aptes à accepter de fortes majorations tarifaires. Ils résilient donc moins.

8. Corrélation entre les variables explicatives deux à deux :

L'étude de corrélation des variables explicatives est une étape primordiale avant toute démarche de modélisation utilisant les modèles linéaires. Elle nous permet d'identifier une suite des variables indépendantes à introduire à notre modèle en évitant la colinéarité afin d'isoler l'effet de chaque variable sur notre variable cible (Y).

Pearson Correlation Coefficients, N = 2737739					
	Puissance_fiscale	Anciennete_contrat	Age_conducteur	age_permis	Age_vehicule
Puissance_fiscale VEPUI	1.00000	0.00493	0.07200	0.07448	0.07747
Anciennete_contrat	0.00493	1.00000	0.23547	0.27691	-0.06075
Age_conducteur	0.07200	0.23547	1.00000	0.65518	-0.07429
age_permis	0.07448	0.27691	0.65518	1.00000	-0.12454
Age_vehicule	0.07747	-0.06075	-0.07429	-0.12454	1.00000

Tableau 2.4 : Matrice de corrélations

On remarque une forte corrélation entre l'âge du conducteur et du permis, ce qui est normal puisque l'ancienneté du permis suppose l'avancement de son titulaire dans l'âge. La variable age_permis sera exclue du vecteur des variables explicatives à introduire lors de la modélisation de rétention.

Chapitre III : Modélisation de la rétention et étude de l'élasticité-prix

Cette section présentera l'approche suivie afin de créer le modèle de rétention ainsi que le bagage théorique associé. On rappelle que notre variable cible est dichotomique, elle prend deux modalités discrètes codées comme suit :

$$Y = \begin{cases} 1 & \text{le client renouvèle sa police} \\ 0 & \text{sinon} \end{cases}$$

Ainsi, le modèle à élaborer estimera la probabilité qu'un assuré renouvellera son contrat. Pour se faire, plusieurs modèles et techniques de prédiction sont envisageables. L'approche classique utilisée dans l'explication du taux de rétention fait appel à une régression logistique de la variable dépendante (binaire) en fonction d'un ensemble varié des variables explicatives. Cette méthode nous permettra d'explicitier analytiquement le lien entre la probabilité de renouvellement et les différents prédicteurs du modèle, on pourra par la suite calculer numériquement l'élasticité de chacun de nos clients par rapport à un changement de prime. Ceci dit, les algorithmes prédictifs de machine learning ne peuvent pas être utilisés dans l'estimation de l'élasticité malgré leur performance comparée au modèle logit.

I. Cadre théorique des Modèles Linéaires Généralisés (GLM) :

1. Formalisation générale du modèle :

Le modèle linéaire généralisé part du même principe que celui du modèle linéaire simple. La différence est qu'au lieu de modéliser la variable à expliquer directement, c'est une fonction de l'espérance de cette variable (appelée fonction de lien) qui est modélisée. La variable à expliquer est supposée suivre d'autres lois que la loi normale. Ces lois font partie de la famille exponentielle qui offre un cadre commun d'estimation et de modélisation.

a. Notations :

Avant de commencer notre formalisation, il convient de spécifier quelques notations et définitions :

- Notant $\mathbf{y} = (Y_1, Y_2, \dots, Y_n)'$ le vecteur des valeurs prises par la variable à expliquer, notée \mathbf{y} .
- Posant $\mathbf{X} = (\mathbf{1}, X_1, X_2, \dots, X_p)'$ où X_j désigne la $j^{\text{ème}}$ variable explicative.

- Soit $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_p)'$ désignant les paramètres du modèle.

b. Composante aléatoire :

Le vecteur \mathbf{y} constitue la partie aléatoire dans ce modèle, il s'agit généralement d'une distribution qui appartient à la famille exponentielle. Rappelant qu'une variable Y a une loi faisant partie de la famille exponentielle si sa densité peut se mettre sous la forme :

$$f(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\Phi}) = \exp\left(\frac{\boldsymbol{\theta}\mathbf{y} - \mathbf{b}(\boldsymbol{\theta})}{\mathbf{a}(\boldsymbol{\Phi})} + \mathbf{c}(\mathbf{y}, \boldsymbol{\Phi})\right)$$

Avec :

- $\boldsymbol{\theta}$: Paramètre réel , aussi appelé paramètre canonique ou encore paramètre de la moyenne.
- $\boldsymbol{\Phi}$: Paramètre réel, appelé paramètre de la dispersion.
- $\mathbf{a}, \mathbf{b}, \mathbf{c}$ sont des fonctions.

c. Composante déterministe :

La composante $\mathbf{X}'\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \dots + \mathbf{X}_p\boldsymbol{\beta}_p$ est la partie déterministe de ce modèle, appelée aussi le prédicteur linéaire. Ainsi, l'idée principale du modèle linéaire généralisé est d'estimer la moyenne $\mu = \mathbf{E}[\mathbf{y}|\mathbf{X}]$ à l'aide du prédicteur $\mathbf{X}'\boldsymbol{\beta}$, en supposant que les Y_i sont indépendantes et associées à une loi de probabilité issue de la famille exponentielle.

d. Fonction de lien :

La relation entre la composante aléatoire et le prédicteur linéaire est exprimée par la troisième composante appelée fonction de lien \mathbf{g} , strictement monotone et dérivable telle que :

$$\mathbf{g}(\mu) = \mathbf{X}'\boldsymbol{\beta}$$

Ainsi, l'espérance conditionnelle de \mathbf{y} correspond à une transformation du prédicteur linéaire. Voici des exemples de fonctions de lien classiques :

Loi	Nom du lien	Fonction de lien
Bernoulli/Binomiale	lien logit	$g(\mu) = \text{logit}(\mu) = \ln(\mu/1 - \mu)$
Normale	lien identité	$g(\mu) = \mu$
Poisson	lien log	$g(\mu) = \ln(\mu)$
Gamma	lien réciproque	$g(\mu) = -1/\mu$

Tableau 3.1 : Fonctions de liens classiques

2. Modèle Logistique :

En utilisant le lien logit, la régression s'appelle alors communément la régression logistique. Cette régression est la plus utilisée lorsqu'il s'agit de modéliser une variable binaire, Autrement dit, de prédire la survenance d'un événement en fonction de certaines caractéristiques. Elle sera notre méthode de référence pour prédire le renouvellement ou la résiliation d'un certain contrat.

$$Y = \begin{cases} \mathbf{1} & \text{avec une probabilité } p \\ \mathbf{0} & \text{avec une probabilité } 1 - p \end{cases}$$

a. Appartenance à la famille exponentielle :

La distribution d'une variable binaire appartient à la famille exponentielle.

Preuve :

$$P(Y = y_i) = p^{y_i} * (1 - p)^{1-y_i} \quad \text{Où } y_i \in \{1, 0\}$$

En Posant : $\theta = \mathbf{Log}\left(\frac{p}{1-p}\right)$, $\mathbf{b}(\theta) = \mathbf{log}(1 + \mathbf{exp}(\theta))$, $\mathbf{a}(\varphi) = \mathbf{1}$, $\mathbf{c}(y, \varphi) = \mathbf{0}$. On montre facilement que la distribution de Y appartient à la famille exponentielle.

b. Estimation de paramètres :

La méthode d'estimation utilisée dans le cadre d'un modèle linéaire généralisé est le maximum de vraisemblance. Tout d'abord, nous considérons un échantillon (y_1, y_2, \dots, y_n) de notre variable Y .

On a pour le modèle logistique :

$$\mu = E[Y|X] = P[Y = 1|X] = g^{-1}(X'\beta) / g(x) = \text{Logit}(x) = \text{Log}\left(\frac{x}{1-x}\right)$$

Ainsi :

$$P[Y = 1|X] = \frac{e^{X'\beta}}{1+e^{X'\beta}} = F(X'\beta)$$

La vraisemblance s'écrit:

$$L = \prod_{i=1}^n P(y_i = 1)^{y_i} * (1 - P(y_i = 1))^{1-y_i}$$

L'expression de la log-vraisemblance :

$$\text{Log}(L) = \sum_{i=1}^n y_i * \text{Log}(F(X_i'\beta)) + (1 - y_i) * (\text{Log}(1 - F(X_i'\beta)))$$

On cherche à maximiser cette dernière expression, ce qui consiste à calculer tout d'abord les dérivées par rapport aux paramètres β_k :

$$\frac{\partial \text{Log}(L)}{\partial \beta_k} = \sum_{i=1}^n \left(y_i * \frac{f(X_i'\beta)}{F(X_i'\beta)} - (1 - y_i) * \frac{f(X_i'\beta)}{1 - F(X_i'\beta)} \right) * X_{ik} ; \quad \forall k \in \{1, \dots, p\}$$

Où :

$$f(x) = F'(x) = F(x) * (1 - F(x))$$

Ainsi, on cherche β vérifiant le système d'équations suivant :

$$\frac{\partial \text{Log}(L)}{\partial \beta_k} = \sum_{i=1}^n \left(y_i * (1 - F(X_i'\beta)) - (1 - y_i) * F(X_i'\beta) \right) * X_{ik} = 0 ; \quad \forall k \in \{0, \dots, p\}$$

La solution analytique de ce système n'existe pas, on utilisera l'algorithme de Newton-Raphson pour s'approcher de β . Pour utiliser cette méthode, nous avons besoin de la dérivée seconde de la log-vraisemblance par rapport β . l'algorithme utilise la relation suivante :

$$\beta^{(i+1)} = \beta^{(i)} - \left(\frac{\partial^2 \text{Log}(L)}{\partial \beta^2} \right)^{-1} * \frac{\partial \text{Log}(L)}{\partial \beta}$$

On fixe β au départ. L'algorithme s'arrête lorsque la différence est suffisamment faible.

Propriétés statistiques de l'estimateur du maximum de vraisemblance :

- ✓ L'estimateur du maximum de vraisemblance $\hat{\beta}$ converge presque sûrement vers la vraie valeur β .
- ✓ L'estimateur du maximum de vraisemblance $\hat{\beta}$ est asymptotiquement sans biais et à variance minimale.
- ✓ L'estimateur du maximum de vraisemblance $\hat{\beta}$ est asymptotiquement normal.

c. Significativité des variables explicatives : Test de type III

Avant d'approfondir la modélisation, il est nécessaire de tester la significativité de chaque variable explicative. Il s'agit de vérifier est ce que la variable retenue contribue à la connaissance de la variable à expliquer. Rappelons que le modèle étudié s'écrit :

$$g(E[y|X]) = X' \beta$$

Hypothèses du test :

$$\begin{cases} H_0: \beta_i = 0 \text{ (la variable } X_i \text{ n'apporte pas d'information sur } y \text{)} \\ H_1: \beta_i \neq 0 \text{ (la variable } X_i \text{ apporte d'information sur } y \text{)} \end{cases}$$

Sous l'hypothèse H_0 , en notant $\hat{\beta}_i$ l'estimateur du maximum de vraisemblance de β dans le modèle privé de la variable X_i , la statistique de test est définie comme suit :

$$S = \frac{2 \left(l(\hat{\beta}) - l(\hat{\beta}_i) \right)}{\Phi}$$

Cette statistique suit une loi du **Khi – deux** à $k - 1$ degrés de liberté où k étant le nombre de modalités de la variable X_i .

Ainsi, La variable n n'est pas significative si $S_{calculée}$ dépasse le quantile d'ordre $1 - \alpha$ de la loi **Khi – deux** à $k - 1$ degrés de liberté.

d. Significativité des coefficients : Test de Wald

À ce niveau on cherche à tester la nullité d'un coefficient pour s'assurer de son significativité. Dans la pratique, ce test permet de vérifier la pertinence des variables, mais aussi de modifier les regroupements de modalité. Si une modalité n'est pas significative, on sera alors amené à agréger les modalités de manière cohérente afin de créer une nouvelle variable significative. Hypothèses du test :

$$\begin{cases} H_0: \beta_i = 0 \text{ (le coefficient } \beta_i \text{ est significativement nulle)} \\ H_1: \beta_i \neq 0 \text{ (le coefficient } \beta_i \text{ est significativement non nulle)} \end{cases}$$

Sous l'hypothèse H_0 , en notant $\hat{\beta}_i$ l'estimateur du maximum de vraisemblance de β_i la statistique de test est définie comme suit :

$$W = \frac{\hat{\beta}_i^2}{v(\hat{\beta}_i)}$$

La statistique suit alors une loi du **Khi – deux** à 1 à degré de liberté. Ainsi, Le coefficient est jugé statistiquement nul si $W_{calculée}$ dépasse le quantile d'ordre $1 - \alpha$ de la loi **Khi – deux** à 1 degrés de liberté.

e. Validation du modèle : Test d'hypothèse global ($\beta = 0$)

Plusieurs tests sont utilisés afin de valider le modèle, dans cette section on présentera le test de rapport de vraisemblance (LR). L'idée de test étant de comparer le modèle complet (selon toutes les variables) avec le modèle nul ($\beta_i = 0 \forall i \in \{1, \dots, p\}$). Hypothèses du test :

$$\begin{cases} H_0: \beta_1 = \beta_2 \dots = \beta_p = 0 \text{ (Aucune des variables explicatives n'influence la variable réponse)} \\ H_1: \exists i \in \{1, \dots, p\} / \beta_i \neq 0 \text{ (Au moins une variable explicative influence la variable réponse)} \end{cases}$$

Sous l'hypothèse H_0 , en notant $\hat{\beta}$ l'estimateur du maximum de vraisemblance de β dans le modèle complet et $\hat{\beta}^c$ l'estimateur dans le modèle nul, la statistique de test est définie comme suit :

$$LR = -2 * (l(\hat{\beta}) - l(\hat{\beta}^c))$$

Cette statistique suit une loi du **Khi – deux** à **p** degrés de liberté. Ainsi, on rejette l’hypothèse nulle. Si $LR_{calculée}$ dépasse le quantile d’ordre $1 - \alpha$ de la loi **Khi – deux** à **p** degrés de liberté.

f. Performance du modèle logistique :

i. Matrice de confusion :

La matrice de confusion montre le nombre des prédictions correctes et incorrectes faites par notre modèle de classification d’une variable binaire, elle compare les valeurs attendues avec celles réellement observées :

Matrice de confusion		Prédit par le modèle			
		1	0		
Observé	1	a	b	Sensibilité	$a/(a+b)$
	0	c	d	Spécificité	$d/(c+d)$
Taux d’erreur = $(c+b)/(a+b+c+d)$		Précision = $a/(a+c)$		Taux de succès = $(a+d)/(a+b+c+d)$	

Tableau 3.2 : Matrice de confusion

- a sont les **Vrais Positifs** (VP) c.-à-d. les observations qui ont été classées positives et qui le sont réellement ;
- c sont les **Faux Positifs** (FP) c.-à-d. les individus classés positifs et qui sont réalité des négatifs ;
- de la même manière, b sont les **Faux Négatifs** (FN) et d sont les **Vrais Négatifs** (VN) ;
- Le **taux d’erreur** est égal au nombre de mauvais classement rapporté à l’effectif total ;
- Le **taux de succès** correspond à la probabilité de bon classement du modèle, c’est le complémentaire à 1 du taux d’erreur ;
- La **sensibilité** (ou le rappel, ou encore le taux de vrais positifs [TVP]) indique la capacité du modèle à retrouver les positifs ;
- La **précision** indique la proportion de vrais positifs parmi les individus qui ont été classés positifs
- La **spécificité**, à l’inverse de la sensibilité, indique la proportion de négatifs détectés.

ii. Courbe ROC :

La courbe ROC est une représentation qui met en relation le Taux de Vrais Positifs (TVP, la sensibilité) et le Taux de Faux Positifs (TFP, 1 - Spécificité) dans un graphique nuage de points. Habituellement, nous comparons $\widehat{\mathbf{p}}(\mathbf{x})$ à un seuil $s = 0.5$ pour effectuer une prédiction du $\mathbf{y}(\mathbf{x})$. Nous pouvons ainsi construire la matrice de confusion et en extraire les 2 indicateurs précités. La courbe ROC généralise cette idée en faisant varier le seuil s sur tout l'intervalle $[0,1]$. Ainsi, Pour chaque seuil s nous construisons la matrice de confusion et nous calculons TVP et TFP.

Le taux de vrais positifs (TVP) indique la capacité du modèle à trouver les positifs :

$$TVP(s) = \frac{\sum_{i=1}^N 1_{\{y_i(x_i)=1\} \cap \{\widehat{p}_i(x_i) \geq s\}}}{\sum_{i=1}^N 1_{\{y_i(x_i)=1\}}}$$

Inversement, le taux de faux positifs, indique la proportion de négatifs détectés :

$$TFP(s) = \frac{\sum_{i=1}^N 1_{\{y_i(x_i)=0\} \cap \{\widehat{p}_i(x_i) \geq s\}}}{\sum_{i=1}^N 1_{\{y_i(x_i)=0\}}}$$

Formellement, la courbe ROC est défini : $\{(TVP(s), TFP(s)) / s \in [0,1]\}$

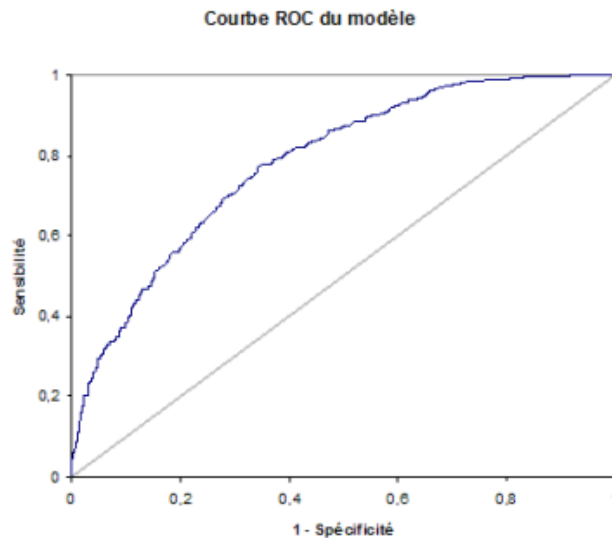


Figure 3.1: Exemple de courbe de ROC

Ainsi, Le meilleur modèle correspond à une courbe qui s'éloigne de la bissectrice, cette mesure est quantifiée à l'aide de la statistique AUC (Area Under the Curve) calculant l'aire

sous la courbe. Elle est comprise entre 0.5 et 1. Plus l'aire est proche de 1, plus la prédiction est bonne.

Valeur de l'AUC	Commentaire
$AUC = 0.5$	Pas de discrimination.
$0.7 \leq AUC < 0.8$	Discrimination acceptable
$0.8 \leq AUC < 0.9$	Discrimination excellente
$AUC \geq 0.9$	Discrimination exceptionnelle

Tableau 3.3 : Interprétation des valeurs du critère AUC

II. Modélisation de la rétention :

1. Formulation du modèle :

On considère la formulation suivante du modèle de rétention :

$$P(Y_i = 1 | \mathbf{X}_i, P_i) = E(Y_i | \mathbf{X}_i, P_i) = \frac{1}{(1 + \exp(-(\beta_0 + \beta_1 * \ln(P_i) + \beta * \mathbf{X}_i)))}$$

Où :

$$\left\{ \begin{array}{l} P \text{ (Premium)} \quad : \text{ prime demandée lors du renouvellement ;} \\ X \quad : \text{ Autres variables explicatives ;} \\ Y \quad : \text{ Variable à expliquer ;} \\ i \quad : \text{ Individu ;} \end{array} \right.$$

Les assureurs cherchent à construire une fonction de demande qui vérifie les propriétés suivantes :

- ✓ La fonction de demande dépend du prix : β_1 significatif ;
- ✓ La probabilité de la demande décroît avec le prix : $\beta_1 < 0$;
- ✓ La probabilité de la demande est nulle pour un prix infini :

$$\lim_{P_i \rightarrow +\infty} P(Y_i = 1 | \mathbf{X}_i, P_i) = 0$$

- ✓ La demande atteint son niveau maximal pour une prime nulle, c'est-à-dire que la probabilité de renouvellement doit tendre vers 1 quand la prime se rapproche de 0. C'est grâce à la transformation log qu'on arrive à vérifier cette propriété :

$$\lim_{P_i \rightarrow 0} P(Y_i = 1 | \mathbf{X}_i, P_i) = 1$$

2. Résultats de la modélisation :

Nous vous présenterons dans cette partie les différents outputs de la modélisation en interprétant l'ensemble des résultats obtenus. Ce point n'est qu'une mise en pratique des différents concepts théoriques introduits précédemment sur notre base de données. On rappelle que l'objectif de cette modélisation se concrétise en quantifiant la probabilité que notre client renouvèle sa police d'assurance en fonction de deux types de paramètres : Marché (concurrence, prix, service, offre,...) et Client (caractéristiques conducteur, sa vie socio-économique, caractéristiques véhicule,...).

a. Echantillonnage :

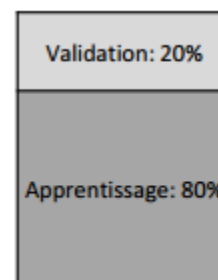
L'échantillonnage constitue une étape fondamentale pour élaborer un modèle prédictif. On subdivise notre population en deux échantillons :

- L'échantillon d'apprentissage : représente généralement 70-80% des données, c'est à partir duquel le modèle est construit. Cette base permet d'estimer les paramètres du modèle.
- L'échantillon de test : cette base sert à confronter les résultats du modèle sur des données neutres et permet d'estimer les erreurs entre les prédictions et les réalisations afin de valider le modèle retenu et affirmer sa justesse.

Cette technique est très utile dans le cas des algorithmes prédictifs de l'apprentissage automatique supervisé, elle nous permet de s'assurer qu'il y a absence du problème de sur-apprentissage : on obtient un modèle sur-ajusté avec un grand nombre de paramètres (cas extrême : assez de paramètres que d'observations), on dit dans ce cas que le modèle mémorise par cœur les données, il perd par la suite son pouvoir de généralisation et sa capacité prédictive si on l'applique sur de nouvelles observations.

Notons bien que ces échantillons doivent être représentatifs pour garantir la qualité du modèle. On doit s'assurer après cet échantillonnage que chaque côté contient une proportion raisonnable des variables explicatives : en faisant l'échantillonnage il se peut qu'une modalité d'une variable catégorielle très intéressante dans notre modèle ne se sélectionne que dans la base test, elle sera totalement absente dans la base apprentissage, on obtiendra par la suite un modèle biaisé.

Nous avons utilisé la procédure « surveyselect » pour découper la base de données en 80% apprentissage - 20% test :



b. Estimation des paramètres :

Le modèle est lancé sur la base apprentissage. Nous avons choisi la méthode «stepwise» pour sélectionner l'ensemble des variables explicatives qui constituent le meilleur modèle parmi les différentes combinaisons possibles tout en minimisant les deux fameux critères d'information AIC et BIC. Le tableau suivant présente quelques paramètres du modèle obtenu :

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	7.3464	0.0952	5952.6195	<.0001
log_Prime		1	-1.0134	0.00771	17272.4983	<.0001
dist_BP		1	0.000044	3.466E-6	161.3096	<.0001
Type_echeance	DC	1	0.3574	0.00632	3199.5071	<.0001
Type_echeance	DF	0	0	.	.	.
Sexe	F	1	0.3905	0.0520	56.3134	<.0001
Sexe	M	0	0	.	.	.
Anciennete_contrat		1	0.1032	0.00103	10089.8901	<.0001
Age_conducteur		1	0.0118	0.000127	8558.6736	<.0001

Tableau 3.4 : Extrait du tableau des paramètres du modèle

On remarque d'une part que le paramètre associé à la variable « dist_BP » -mesurant l'écart à la concurrence- a un signe positif, ce qui est illogique puisque plus qu'on demande une prime plus chère (loin du meilleur tarif sur le marché), la probabilité que le client (qui a une bonne visibilité sur les prix du marché) résilie augmente. D'autre part, ce coefficient est pratiquement nul même s'il sort significatif, donc l'effet de cette variable sur notre probabilité sera très faible. Ce problème est dû principalement au benchmark utilisé dont les variables prennent des valeurs discrètes, il fallait donc segmenter les variables de la base de données communs avec ce benchmark pour faire la jointure entre les deux, ce qui nous mène à une variable non fiable qu'on n'en peut pas faire confiance. On décide alors de retirer cette variable de notre modèle finale malgré son importance théorique dans l'explication de la rétention.

Les résultats du modèle finale retenu se présentent comme suit :

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	7.1310	0.0936	5799.4657	<.0001
log_Prime		1	-0.9822	0.00747	17279.3285	<.0001
Type_echeance	DC	1	0.3664	0.00617	3530.4358	<.0001
Type_echeance	DF	0	0	.	.	.
Sexe	F	1	0.3725	0.0514	52.4369	<.0001
Sexe	M	0	0	.	.	.
Anciennete_contrat		1	0.1018	0.00101	10125.7839	<.0001
Age_conducteur		1	0.0117	0.000126	8551.5097	<.0001
Type_garantie	DC	1	0.1369	0.00632	469.4379	<.0001
Type_garantie	DTA	1	0.4041	0.00917	1943.9310	<.0001
Type_garantie	RC	0	0	.	.	.
Energie	E	1	-0.3424	0.00439	6094.9384	<.0001
Energie	G	0	0	.	.	.
Puissance_fiscale		1	0.0577	0.00114	2571.8455	<.0001
taux_CRM	90%	1	0.4177	0.00445	8792.2610	<.0001
taux_CRM	>100%	1	0.2254	0.0102	487.9915	<.0001
taux_CRM	100%	0	0	.	.	.
Taux_intermediaires		1	1.3940	0.0555	630.2235	<.0001
Taux_activite		1	0.4535	0.1040	19.0111	<.0001
Taux_chomage		1	-0.7428	0.0918	65.5113	<.0001
Region	C1	1	0.0565	0.0103	30.2978	<.0001
Region	C2	1	0.2237	0.00992	508.1449	<.0001
Region	C3	1	-0.0231	0.00885	6.8146	0.0090
Region	Chaouia-Ouadigha	1	-0.0602	0.0119	25.8249	<.0001
Region	Doukkala-Abda	1	-0.1989	0.0121	268.1271	<.0001
Region	Grand Casablanca	1	-0.3831	0.0117	1063.7756	<.0001
Region	Meknès-Tafilalet	1	0.0593	0.0112	27.8125	<.0001
Region	Régions de la Sahara	1	-0.5701	0.0219	676.1330	<.0001
Region	Souss-Massa-Drâa	1	0.0318	0.0115	7.5982	0.0058
Region	Tadla-Azilal	1	0.1089	0.0148	53.9609	<.0001
Region	Marrakech-Tensift-Al Haouz	0	0	.	.	.

Tableau 3.5 : Paramètres estimés du modèle retenu

Toutes les variables du modèle sont significatives, on rejette clairement l'hypothèse H_0 : « $\beta_i = 0$ » au seuil de 5% (test de Wald).

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
log_Prime	0.375	0.369	0.380
Type_echeance DC vs DF	1.442	1.425	1.460
Sexe F vs M	1.451	1.312	1.605
Anciennete_contrat	1.107	1.105	1.109
Age_conducteur	1.012	1.011	1.012
Type_garantie DC vs RC	1.147	1.133	1.161
Type_garantie DTA vs RC	1.498	1.471	1.525
Energie E vs G	0.710	0.704	0.716
Puissance_fiscale	1.059	1.057	1.062
taux_CRM 90% vs 100%	1.518	1.505	1.532
taux_CRM >100% vs 100%	1.253	1.228	1.278
Taux_intermediaires	4.031	3.615	4.494
Taux_activite	1.574	1.284	1.930
Taux_chomage	0.476	0.397	0.570
Region C1 vs Marrakech-Tensift-AI Haouz	1.058	1.037	1.080
Region C2 vs Marrakech-Tensift-AI Haouz	1.251	1.227	1.275
Region C3 vs Marrakech-Tensift-AI Haouz	0.977	0.960	0.994
Region Chaouia-Ouardigha vs Marrakech-Tensift-AI Haouz	0.942	0.920	0.964
Region Doukkala-Abda vs Marrakech-Tensift-AI Haouz	0.820	0.800	0.839
Region Grand Casablanca vs Marrakech-Tensift-AI Haouz	0.682	0.666	0.698
Region Meknès-Tafilalet vs Marrakech-Tensift-AI Haouz	1.061	1.038	1.085
Region Régions de la Sahara vs Marrakech-Tensift-AI Haouz	0.565	0.542	0.590
Region Souss-Massa-Drâa vs Marrakech-Tensift-AI Haouz	1.032	1.009	1.056
Region Tadla-Azilal vs Marrakech-Tensift-AI Haouz	1.115	1.083	1.148

Tableau 3.6 : Estimations des rapports de cotes

Les deux tableaux ci-dessus résument l'effet de chaque variable sur la probabilité de renouvellement d'un client, on peut s'en sortir des conclusions intéressantes :

- ✓ $\beta_1 \approx -1 < 0$: une hausse de la prime demandée au client entrainera une baisse de sa probabilité de renouvellement. Plus précisément, si on multiplie la prime par e ($=2.72$) l'odds de renouvellement de l'assuré baissera de 62.5% ;
- ✓ Les clients ayant choisi de souscrire un contrat qui se renouvèle automatiquement (Type échéance = 'DC'), leur chance de ne pas résilier augmente de 44% ;

- ✓ Les femmes sont relativement plus fidèles que les hommes ($\beta > 0$) ;
- ✓ L'ancienneté du contrat à un effet positif sur la rétention : les clients les plus anciens ont moins de chance de résilier ;
- ✓ Nos assurés les plus âgés ont un taux de rétention significativement supérieur à celui des jeunes ;
- ✓ Les renouvellements sont plus fréquents chez les clients ayant souscrits DTA ou DC : les gens qui détiennent la garantie DTA (resp DC) ont 50% (resp 15%) plus de chance de renouveler leurs contrats ;
- ✓ L'effet de la variable CRM qui nous renseigne sur la sinistralité du client se résume comme suit :

$P(Y_i = 1 | CRM = 0.9) > P(Y_i = 1 | CRM > 1) > P(Y_i = 1 | CRM = 1)$ (Toute autre variable explicative étant égale par ailleurs) Nous avons déjà rencontré et interprété ce résultat dans la partie exploratoire de données (voir figure 2.11).

- ✓ Les assurés qui conduisent un véhicule essence ont 29% moins de chance de renouveler que les clients ayant un véhicule gasoil ;
- ✓ La variable puissance impact positivement la rétention ;
- ✓ La région « Marrakech-Tensift-Al Haouz » a été choisie comme modalité de référence. Les autres régions peuvent être classées en deux groupes :

Les régions ($\beta > 0$) : Rabat-Salé-Zemmour-Zaër (C1), Taza-Al Hoceïma-Taounate (C2), L'Oriental (C2), Gharb-Chrarda-Beni Hssen (C2), Meknès-Tafilalet, Souss-Massa-Drâa et Tadla-Azilal. Ces régions connaîtront différents taux de rétention supérieurs à celui de la région de référence ;

Les régions ($\beta < 0$): Grand Casablanca, Tanger-Tétouan (C3), Fès-Boulemane (C3), Chaouia-Ouardigha, Doukkala-Abda et les régions de la Sahara. Les résidents de ces régions ont des probabilités de renouvellement inférieures.

- ✓ Taux des intermédiaires par région : c'est le rapport entre le nombre des intermédiaires d'AAM commercialisant les garanties auto sur l'ensemble des intermédiaires dans chaque région. Cette variable se considère comme indicateur de

concurrence, son coefficient associé est positif : En renforçant son réseau de distribution, AAM réalisera des niveaux de rétention élevés ;

- ✓ Quand le taux d'activité (respectivement chômage) augmente, le taux de rétention augmente (respectivement baisse).

Toutes ces résultats sont conforme avec les conclusions retenues dans la partie « analyse exploratoire des données » du chapitre précédent.

c. Evaluation du modèle :

i. Test globale de significativité :

On rappelle les deux hypothèses du test :

$H_0 : \langle \beta_1 = \beta_2 = \dots = \beta_p = 0 \rangle \Leftrightarrow \langle \text{Pas de liaison entre } Y \text{ et les } X_j \rangle$

$H_1 : \langle \exists \beta_j \neq 0 \rangle \Leftrightarrow \langle \text{Il existe au moins une variable } X_j \text{ expliquant } Y \rangle$

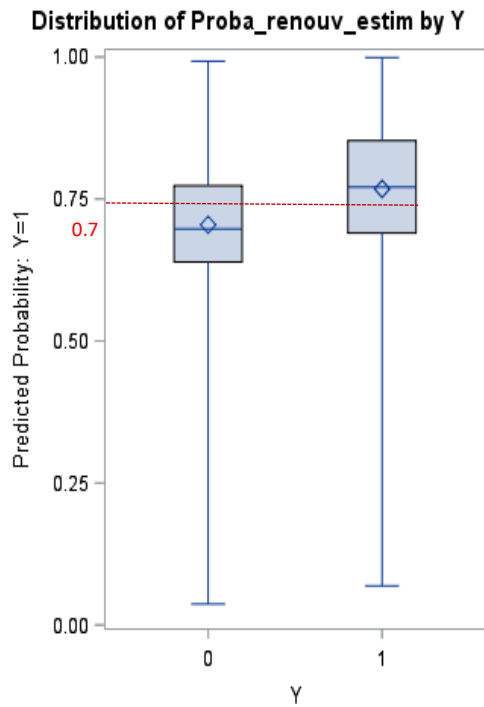
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	139025.307	22	<.0001
Score	130013.955	22	<.0001
Wald	119882.701	22	<.0001

Tableau 3.7 : Tests globaux de significativité du modèle

Les trois tests globaux de la régression présentés dans cette sortie confirment que le modèle retenu est significatif.

ii. Matrice de confusion :

La matrice de confusion est utilisée pour évaluer la capacité du modèle à bien classer les observations (renouveau / résiliation). Elle confronte les valeurs observées de la variable dépendante avec celles qui sont prédites, puis comptabilise les bonnes et les mauvaises prédictions. A partir d'un seuil donné de probabilité, on peut prédire la décision du client. Ce seuil doit être choisi avec prudence en minimisant le nombre des observations mal classifiées. Les figures ci-après montrent la distribution de la probabilité de renouvellement estimée sur la base test. L'intérêt d'utiliser cette base réside dans le fait qu'elle contient de nouvelles observations que notre modèle croise pour la première fois :



Quantiles (Definition 5)	
Level	Quantile
100% Max	0.992713
99%	0.932442
95%	0.886956
90%	0.853483
75% Q3	0.773315
50% Median	0.697379
25% Q1	0.639077
10%	0.572227
5%	0.528460
1%	0.462042
0% Min	0.036966

« Y=0 »

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.9990837
99%	0.9482958
95%	0.9195922
90%	0.8980735
75% Q3	0.8529049
50% Median	0.7709548
25% Q1	0.6898902
10%	0.6363045
5%	0.6008688
1%	0.5203872
0% Min	0.0690032

« Y=1 »

Figure 3.2 : Distribution de la probabilité de renouvellement estimée

Si on prend 0.5 comme seuil de probabilité pour distinguer le renouvellement de la résiliation c-à-d :

$$\begin{cases} P > 0.5 \Rightarrow \hat{Y} = 1 & \text{"renouvellement"} \\ P \leq 0.5 \Rightarrow \hat{Y} = 0 & \text{"résiliation"} \end{cases}$$

Dans ce cas, plus que 95% des résiliations seront prédites comme renouvellement.

En fixe 0.7 comme seuil de décision pour classifier les observations : presque la moitié des résiliations et ¾ des renouvellements seront bien classifiées.

Frequency Percent Row Pct Col Pct	Table of Y by Y_pred		
	Y	Y_pred	
		0	1
0	69732 12.74 51.30 37.24	66201 12.09 48.70 18.37	135933 24.83
1	117502 21.46 28.55 62.76	294114 53.71 71.45 81.63	411616 75.17
Total	187234 34.19	360315 65.81	547549 100.00

Tableau 3.8 : Matrice de confusion

Taux d'erreur = 33.5%

Taux de succès = 66.5%

} $\frac{2}{3}$ des observations sont prédites correctement.

Sensibilité (rappel) = 71.45% des renouvellements sont bien classés.

Précision = 81.63% des observations scorées comme renouvellement le sont effectivement.

Spécificité = 51.3% : notre modèle réussi à prédire la moitié des résiliations. Si on augmente d'avantage le seuil de probabilité fixé, ce taux augmentera au détriment du taux de sensibilité.

iii. Courbe ROC :

La courbe ROC est un outil très utilisé pour évaluer la performance des classificateurs binaires, elle croise dans un plan la sensibilité (ordonnée) et 1-spécificité (abscisse) calculées en variant le seuil de probabilité entre 0 et 1. Les deux graphiques ci-après tracent l'allure de cette courbe en exploitant les résultats du modèle sur les bases apprentissage et test :

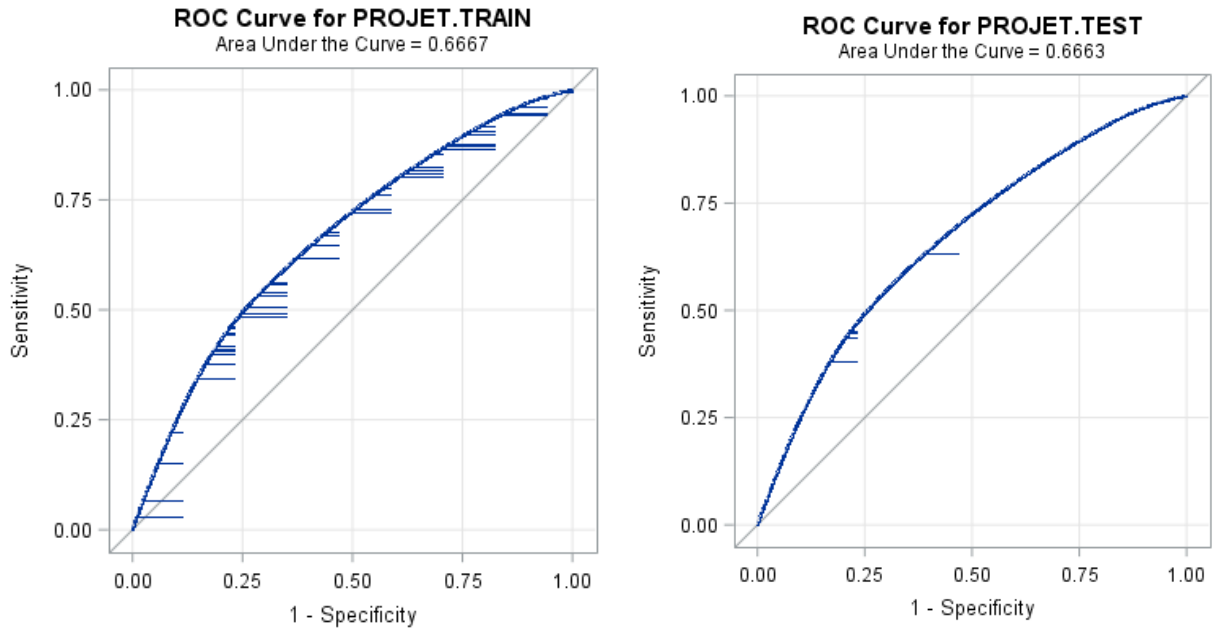


Figure 3.3: Courbes ROC pour les bases apprentissage et test

On remarque que ($AUC_{\text{apprentissage}} \approx AUC_{\text{test}}$), alors le modèle est robuste car son pouvoir prédictif n'a pas changé.

D'après la valeur de $AUC = 0.67$, On déduit que le pouvoir discriminant du modèle n'est pas si bon mais il reste acceptable. Les variables utilisées dans la tarification ne suffisent pas pour élaborer un modèle de rétention performant, il manque des variables reflétant la compétitivité des assureurs et la situation socio-économique de l'assuré.

Nous avons essayé de construire d'autres classificateurs binaires dans le cadre des modèles prédictifs du machine learning disponibles sur SAS Enterprise Miner.

Noeud du modèle	Description du modèle	Critère de sélection : Valid: Misclassification Rate	Test: Misclassification Rate	Apprentissage : Index Roc	Apprentissage : Coefficient de Gini	Test : Index Roc	Test : Coefficient de Gini
HPDMForest	Forest HP	0.2258092886	0.2265207744	0.723	0.445	0.716	0.432
Tree	Arbre de décision	0.2278210587	0.2285746742	0.671	0.343	0.67	0.34
HPNNA	Réseau neuronal HP	0.2282094879	0.2286278081	0.701	0.402	0.699	0.398
HPSVM	SVM HP	0.2377479901	0.2380819764	0.632	0.265	0.631	0.261

Tableau 3.9 : Comparaison entre les algorithmes de prédiction

Le Random forest-RI & les réseaux neurones nous ont permis d'atteindre le niveau de 70% pour la statistique AUC.

Ainsi on observe que :

- Les taux d'erreur (Misclassification Rate) sont très proches entre eux pour les algorithmes en question (**Taux d'erreur \approx 23%, Taux de succès \approx 77%**).
- $AUC_{\text{apprentissage}} \approx AUC_{\text{test}}$ pour chaque algorithme.

Le Random Forest-RI étant le plus performant (AUC=72%). La matrice de confusion associée (Seuil = 0.7) se présente comme le suivant :

Frequency Percent Row Pct Col Pct	Table of y by l_y			
	y	l_y(Into: y)		
		0	1	Total
0	70549 12.93 52.60 43.10	63586 11.65 47.40 16.64	134135 24.58	
1	93126 17.06 22.62 56.90	318527 58.36 77.38 83.36	411653 75.42	
Total	163675 29.99	382113 70.01	545788 100.00	

Tableau 3.10 : Matrice de confusion

Taux d'erreur = 28.7% }
Taux de succès = 71.3% } 70% des observations sont prédites correctement.

Sensibilité (rappel) = 77,38% des renouvellements sont bien classés.

Précision = 83.36% des observations scorées comme renouvellement le sont effectivement.

Spécificité = 52,6% : le modèle réussi à prédire 52,6% des « non-renouvellements ».

On observe une amélioration remarquable des indicateurs de performance par rapport au modèle logistique.

Les autres sorties ainsi qu'un cadre théorique expliquant l'algorithme Random Forest-RI se trouvent dans l'annexe I.

III. Calcul d'élasticité-prix :

Quantifier l'élasticité-prix est l'un des objectifs principaux derrière toute démarche de modélisation de la demande d'un bien déterminé. Un modèle de rétention sophistiqué est important pour valoriser la sensibilité des clients aux prix de la manière la plus réaliste possible. En effet, sous-estimer ou surestimer l'élasticité-prix conduira l'assureur à prendre des décisions injustes qui vont affecter non seulement la rentabilité du portefeuille, mais aussi la loyauté de certains clients et l'image de la compagnie.

1. Cadre théorique :

a. Rappel :

Elasticité-prix de la demande est une mesure de la sensibilité d'un consommateur suite à un changement de prix . On la définit comme suit :

$$Elasticité = - \frac{\frac{dD}{D}}{\frac{dP}{P}}$$

Le signe (-) est utilisé pour obtenir des valeurs positives de l'élasticité-prix puisque la demande baisse quand le prix augmente dans le cadre d'un marché compétitif.

b. Elasticité-prix selon le modèle de rétention :

Un modèle de rétention robuste qui capture efficacement l'élasticité prix des clients de la compagnie est essentiel dans la phase de l'optimisation des tarifs, il sera utile pour formuler les bonnes stratégies de tarification afin de maximiser la marge de l'assureur tout en fidélisant ses clients. Ci-dessous une reformulation du modèle de rétention définit dans l'axe précédant :

$$P(Y_i = 1 | \mathbf{X}_i, P_i) = f(P_i, \mathbf{X}_i) = \frac{\mathbf{1}}{1 + \alpha_i \exp(-\beta_1 \ln(P_i))}$$

$$\text{avec } \alpha_i = \exp(-\beta_0 - \beta * \mathbf{X}_i) > 0 \text{ et } \beta_1 < 0$$

On exprime alors l'élasticité-prix de la façon suivante :

$$e(P_i, \mathbf{X}_i) = -\frac{\frac{df(P_i, \mathbf{X}_i)}{f(P_i, \mathbf{X}_i)}}{\frac{dP_i}{P_i}} = -\frac{\partial f(P_i, \mathbf{X}_i)}{\partial P_i} \frac{P_i}{f(P_i, \mathbf{X}_i)} = -\alpha_i \beta_1 P_i^{-\beta_1} f(P_i, \mathbf{X}_i) > 0$$

On montre facilement que l'élasticité est une fonction croissante de la prime, on a asymptotiquement :

$$\lim_{P_i \rightarrow 0} e_i(P_i) = 0$$

$$\lim_{P_i \rightarrow +\infty} e_i(P_i) = -\beta_1$$

L'élasticité-prix atteint une limite supérieure ($-\beta_1$) lorsque le prix tend vers l'infini. On conclut que l'élasticité-prix peut être sous-évaluée pour des grands montants de prime. Ainsi on rencontre un inconvénient de la formulation utilisée pour définir le modèle de rétention. Ce problème peut être pallié en appliquant d'autres fonctions de transformation sur la prime.

Chaque assureur cherche le vecteur des primes qui maximise la marge globale de son portefeuille. En disposant de la probabilité de renouvellement de chaque assuré, on formule un programme d'optimisation simplifié (sans contraintes) relatif aux clients de la compagnie, ainsi le programme suivant négligera la marge due au taux de non-conversion des affaires nouvelles :

$$\max_{P_1, P_2, \dots, P_n} \text{Marge} = \sum_{i=1}^n (P_i - AC_i) * f(P_i, \mathbf{X}_i)$$

Avec : AC_i (Actuarial Cost) la prime déduite du modèle technique. Elle quantifie le risque que présente chaque assuré.

$$\frac{\partial \text{Marge}}{\partial P_i} = f(P_i, \mathbf{X}_i)(1 - e_i(P_i)) + \frac{AC_i}{P_i} e_i(P_i)$$

Alors :

$$e_i(P_i) < \frac{P_i}{P_i - AC_i} \Leftrightarrow \frac{\partial \text{Marge}}{\partial P_i} > 0$$

On déduit que si l'assureur demande à chaque client un prix légèrement supérieur à la prime technique, une forte probabilité que l'élasticité-prix soit inférieure au rapport $\frac{P}{P-AC}$, et par la suite la marge augmentera avec cette augmentation du prix.

2. Résultats de l'élasticité-prix :

a. Elasticité-prix en fonction de la probabilité de renouvellement :

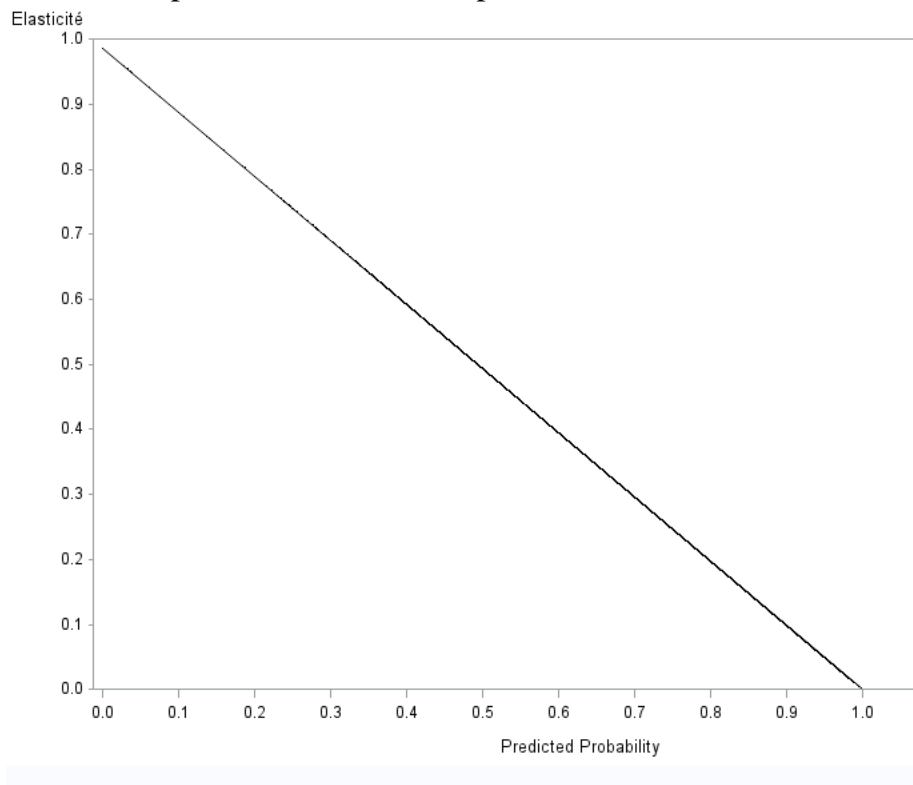


Figure 3.4 : Elasticité-prix en fonction de la probabilité de renouvellement

On constate d'après la figure ci-après que l'élasticité-prix est une fonction décroissante de la probabilité de renouvellement, Ce résultat est logique puisque pour les segments de client ayant une forte probabilité de rétention, On observe une réaction faible par rapport au changement de prix, tandis que cette réaction devient importante pour les segments ayant une faible probabilité de rétention (segments risqués). On dit alors que les segments risqués sont plus élastiques.

Aussi, on remarque que la relation est linéaire, ceci se justifie par la formulation que nous avons considérée.

Preuve :

On rappelle que :

$$\begin{cases} P(Y_i = 1|P_i, \mathbf{X}_i) = f(P_i, \mathbf{X}_i) = \frac{1}{1 + \alpha_i P_i^{-\beta_1}} & (1) \\ e(P_i, \mathbf{X}_i) = -\alpha_i \beta_1 P_i^{-\beta_1} f(P_i, \mathbf{X}_i) & (2) \end{cases}$$

(1) $\Rightarrow \alpha_i = \left(\frac{1}{f(P_i, \mathbf{X}_i)} - 1 \right) * P_i^{\beta_1}$ puis on remplace α_i dans la relation (2), On déduit que l'élasticité s'écrit :

$$e(P_i, \mathbf{X}_i) = -\beta_1 + \beta_1 * P(Y_i = 1|P_i, \mathbf{X}_i) = -\beta_1 * P(Y_i = 0|P_i, \mathbf{X}_i)$$

Cette relation nous permet de projeter et inverser toutes les conclusions sur la rétention déduites dans la partie de la modélisation.

b. Elasticité-prix en fonction de la prime :

Le graphique suivant représente la probabilité de rétention en fonction de la prime au niveau de 6 segments de clients :

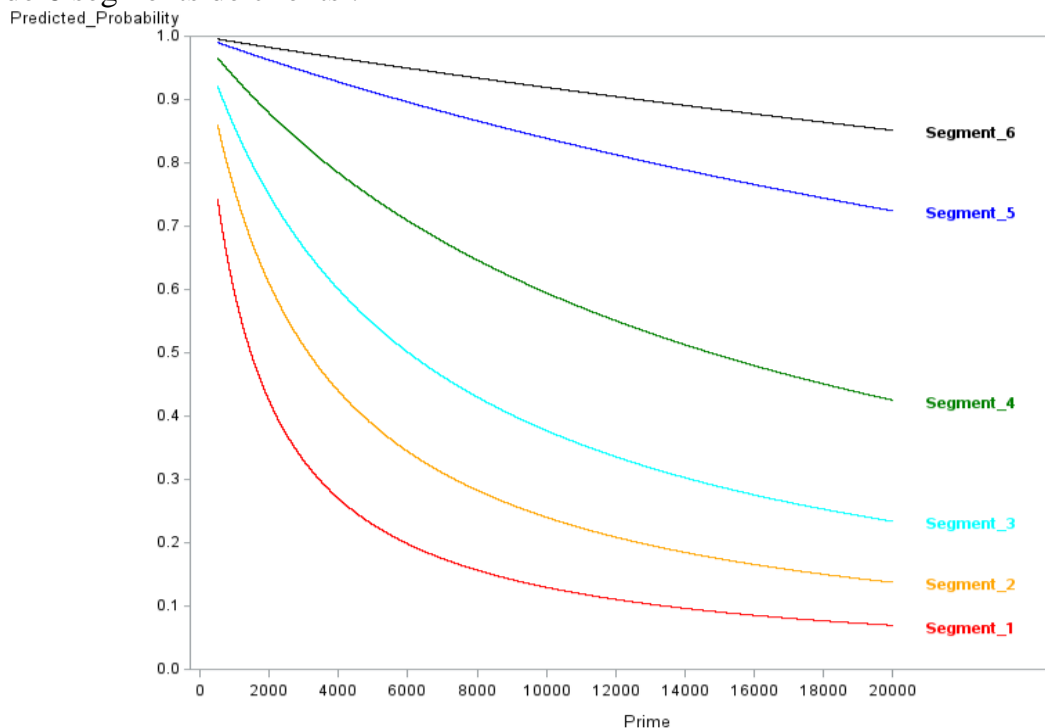


Figure 3.5 : Probabilité de rétention selon différents profils de risque

Segment_1 est un exemple de segment risqué, il regroupe des clients ayant la combinaison des paramètres (autre que β_1) qui augmentent α : Pour les variables catégorielles nous avons sélectionné la modalité ayant le β le plus faible et pour les variables continues nous avons choisi l'une des deux bornes de la variable selon le signe de β . Ces clients se caractérisent par des niveaux de probabilité de renouvellement faibles pour chaque valeur de prime.

Segment_6 est un exemple de segment moins risqué, il regroupe les assurés fidèles ayant le α faible. Le reste des segments ciblent des profils de risque moyen.

On constate que pour le segment_1 et le segment_2, la probabilité de rétention s'approche rapidement vers 0 par rapport autres segments. En effet, ces courbes sont associées à des segments risqués, caractérisés par une réaction élevée face au changement de la prime.

Par ailleurs, nous pouvons distinguer 3 phases importantes marquant l'évolution de la probabilité de rétention pour le segment risqué :

- Pour les niveaux de prime inférieurs, la probabilité de rétention diminue rapidement.
- Pour les valeurs moyennes de la prime, la probabilité de rétention diminue d'une manière moins rapide par rapport au premier cas.
- Pour des valeurs élevées de la prime, la probabilité de rétention diminue très lentement en fonction de l'évolution de prime.

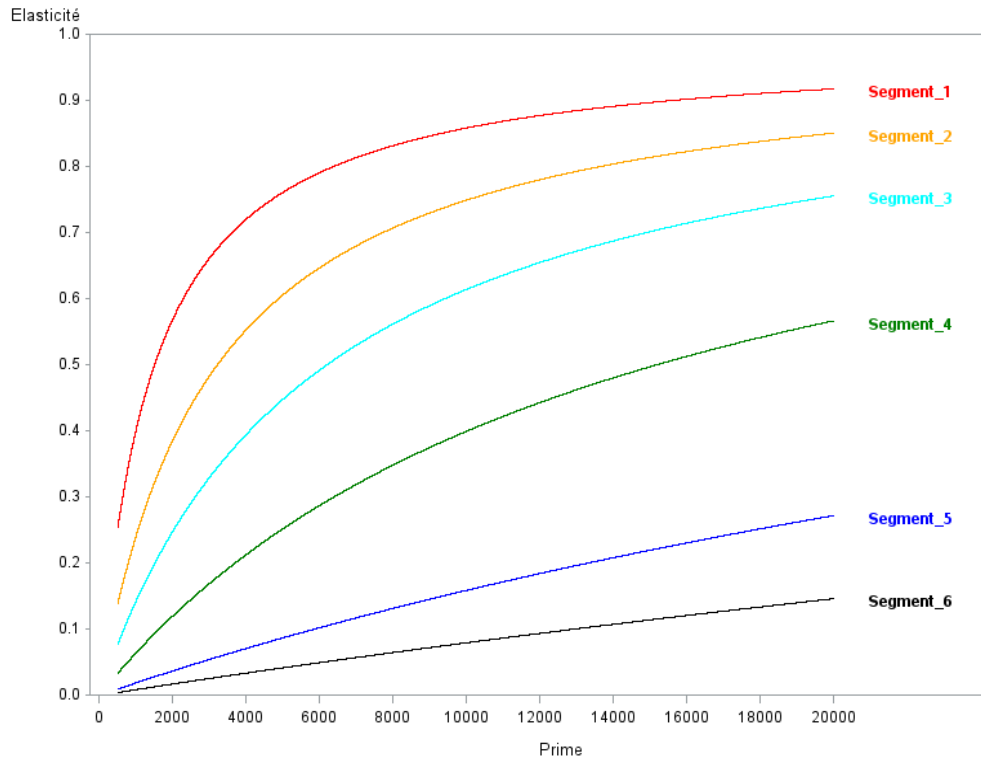


Figure 3.6 : Elasticité-prix selon différents profils de risque

Dans ce graphique, on représente l'évolution de l'élasticité-prix en fonction de la prime pour les segments précités, on remarque que l'élasticité-prix est une fonction croissante de la prime.

En effet :

$$e(P_i, \mathbf{X}_i) = -\beta_1 + \beta_1 f(P_i, \mathbf{X}_i)$$

Alors :

$$\frac{\partial e(P_i, \mathbf{X}_i)}{\partial P_i} = \beta_1 \frac{\partial f(P_i, \mathbf{X}_i)}{\partial P_i}$$

Cette quantité est positive puisque la demande est une fonction décroissante de la prime et $\beta_1 < 0$.

On remarque que l'élasticité des segments risqués (Segment_1 & Segment_2) convergent plus vite à $-\beta_1$ que dans les segments moins risqués.

3. Validation empirique des résultats « Price testing » :

Le « Price testing » introduit dans le premier chapitre de ce rapport, est une technique importante pour valider empiriquement les résultats et les conclusions tirées de la formule analytique de l'élasticité. On évoque ci-après son principe :

L'idée consiste à appliquer différentes variations de prix (positives et négatives [Voir figure 1.9]) sur un échantillon de clients, puis observer comment ils vont réagir vis-à-vis ce changement de tarifs. Ceci nous permet d'explicitier l'élasticité-prix empirique qu'on pourra par la suite la comparer avec l'élasticité-prix analytique déduite du modèle de rétention.

IV. Aperçu sur l'optimisation :

Nous vous introduisons dans cette partie la démarche d'optimisation des tarifs. C'est une consolidation des résultats du modèle de la demande (prédit le comportement [Conversion ou Rétention] de l'assuré) et le modèle technique de la tarification (quantifie le risque assuré).



Figure 3.7 : Aperçu sur l'optimisation

Il existe plusieurs méthodes d'optimisation qui peuvent être utilisées, ils se diffèrent au niveau de la formulation du programme d'optimisation et sa méthode de résolution. On présente ci-après un aperçu sur l'optimisation individuelle, la méthode la plus recommandée dans le cas d'un problème de rétention :

Optimisation individuelle est une méthode permettant d'optimiser les tarifs d'assurance à partir des caractéristiques individuelles de chaque client dans le portefeuille concerné. L'avantage principal de cette technique consiste à estimer la prime optimale à demander à chaque assuré s'il revient pour renouveler son contrat en tenant compte son élasticité-prix et la compétitivité présente sur le marché. Soit la formulation suivante du programme d'optimisation pour un portefeuille de n clients assurés :

$$(P): \begin{cases} \max_{P_1, P_2, \dots, P_n} \sum_{i=1}^n (P_i - AC_i) * f(P_i, \mathbf{X}_i) \\ P_i \leq \overline{P}_i \quad \forall i \in \{1, 2, \dots, n\} & (1) \\ P_i \geq \underline{P}_i \quad \forall i \in \{1, 2, \dots, n\} & (2) \\ \frac{1}{n} \sum_{i=1}^n f(P_i, \mathbf{X}_i) \geq \gamma & (3) \end{cases}$$

Avec :

- P_i : Prime optimale associée au client i ;
- AC_i : Prime technique associée au client i ;
- $f(P_i, \mathbf{X}_i)$: Probabilité de renouvellement du client i estimée à partir du modèle de rétention ;
- \underline{P}_i et \overline{P}_i deux niveaux respectivement inférieur et supérieur de la prime fixés pour le client i ;
- γ : niveau minimum de rétention ciblé.

Fonction objective de ce programme s'interprète comme la marge de rétention espérée suite à une variation de la prime par rapport à celle du modèle technique. Intuitivement $P_i \geq AC_i \quad \forall i \in \{1, 2, \dots, n\}$, l'assureur demandera une prime supérieure au tarif technique pour les clients ayant une élasticité-prix faible et il peut se contenter de la prime technique pour les assurés ayant une faible probabilité de renouveler leurs contrats.

On distingue deux types de contraintes :

- ✓ Contraintes globales : ce type de contraintes regroupe l'ensemble des clients dans le portefeuille (inégalité (3) du programme ci-dessus) ;

- ✓ Contraintes locales : concernent chacune des primes individuelles (contraintes (1) et (2)).

La contrainte globale du programme (P) assure une évolution bien définie de rétention dans le temps conformément à la stratégie de l'entreprise et ses valeurs. Les différentes primes outputs de ce programme doivent vérifier cette contrainte, autrement dit on cherche les tarifs à offrir à chaque client assurant un niveau minimum de rétention dans notre portefeuille : la probabilité moyenne de renouvellement soit supérieure à γ .

Les deux contraintes locales (1) et (2) permettent d'éviter une disparité des prix optimisés en fixant pour chaque individu un intervalle dans lequel on optimise sa prime. Si on ignore ces deux contraintes, on peut tomber sur des primes largement différentes pour des clients de profils proches. Les deux limites de primes peuvent être définies en fonction de la sinistralité de client et les précédentes primes qu'il a payé auparavant, elles garantissent aussi une évolution stable de la prime dans le temps.

L'optimisation des primes est très sensible aux différents seuils considérés dans la formulation des contraintes du programme. Soit un client (i) ayant une faible élasticité-prix, choisir une valeur petite de \bar{P}_i qui apparaît dans la contrainte (1) impliquera une sous-optimisation de la prime. On note aussi que la convergence du programme (P) dépend des trois contraintes.

Etudier d'existence d'une solution au programme (P) n'est pas évident. On peut montrer que la fonction objective du programme (P) admet une solution sous la contrainte globale (3) (Voir annexe II)

Conclusion

Le rôle principal de l'assurance est de mutualiser le risque, offrant ainsi aux assurés une stabilité financière contre les aléas de la vie. Néanmoins, dans un marché libre et concurrentiel, les entreprises sont tenues de créer de la valeur, et de réaliser des bénéfices qui répondent aux exigences de leurs actionnaires. Les sociétés d'assurance n'échappent pas à cette logique des marchés.

Dans ce contexte, les assureurs doivent mieux appréhender la typologie de leurs clients en terme de probabilité de rétention et d'élasticité-prix. L'objectif de notre étude était de proposer une méthodologie de tarification qui prenne en considération ces critères.

Dans un premier temps, nous avons présenté un modèle de régression logistique qui nous a permis d'expliquer la probabilité individuelle de renouvellement en fonction de plusieurs variables explicatives, dont notamment la prime. Nous avons ensuite étudié l'élasticité-prix qui dérive du modèle de rétention. A la fin du dernier chapitre, nous avons présenté un programme d'optimisation qui met en évidence l'utilité de ce genre de modèles de demande en ajustant les résultats des modèles de tarification afin d'accélérer la croissance du portefeuille.

Références

Bibliographie:

- ✓ Documentation d'AAM sur le modèle de la rétention
- ✓ Modèle linéaire généralisé (GLM) . Abdelaziz Chaoubi, INSEA
- ✓ Cours d'économétrie II. Touhami Abdelkhalek, INSEA
- ✓ Économétrie & Machine Learning. Arthur Charpentier, Emmanuel Flachaire, Antoine Ly
- ✓ Arbres CART et Forêts aléatoires, Importance et sélection de variables. Robin Genuer, Jean-Michel Poggi
- ✓ Pratique de la Régression Logistique. Ricco Rakotomalala

Webographie :

<https://www.axa.ma/>

<https://www.fmsar.org.ma/>

<https://www.acaps.ma/>

<https://documentation.sas.com>

<https://www.hcp.ma/>

Annexes

Annexe I : Généralités sur les *forêts* aléatoires

L'acronyme CART (Classification And Regression Trees) désigne une méthode statistique, introduite par Breiman et al. (1984) qui construit des prédicteurs par arbre aussi bien en régression qu'en classification. Le principe général de CART est de partitionner récursivement l'espace d'entrée \mathbf{X} de façon binaire, puis de déterminer une sous-partition optimale pour la prédiction. A l'issue de l'algorithme, on obtient donc un arbre prédisant la variable réponse \mathbf{Y} en fonction d'entrée \mathbf{X} .

Détaillons la règle de construction de l'arbre dite « optimale ». Pour fixer les idées, le lecteur peut se restreindre à des variables explicatives continues (le cas qualitatif se traite de la même manière). L'espace d'entrée noté \mathbf{R}^p , où p est le nombre de variables explicatives. La racine de l'arbre est l'ensemble contenant toutes les observations de l'échantillon apprentissage noté L_n .

On définit la coupure (découpe ou split) un élément de la forme :

$$\{\mathbf{X}_i \leq \mathbf{d}\} \cup \{\mathbf{X}_i > \mathbf{d}\}$$

En partant, de la racine de l'arbre. La première étape de l'algorithme consiste à découper au mieux cette racine en deux nœuds fils. Découper suivant la coupure $\{\mathbf{X}_i \leq \mathbf{d}\} \cup \{\mathbf{X}_i > \mathbf{d}\}$ signifie que toutes les observations avec une valeur de la variable \mathbf{X}_i inférieur à \mathbf{d} vont dans le nœud fils de gauche, et toutes celles avec une valeur supérieure à \mathbf{d} vont dans le nœud fils de droite. La méthode sélectionne alors la meilleure découpe, c'est-à-dire le couple (\mathbf{i}, \mathbf{d}) qui minimise une certaine fonction de coût, l'ensemble des couples (\mathbf{i}, \mathbf{d}) est fini, car le nombre de variables et de données est fini.

Dans le cas de la régression, on cherche à minimiser la variance intra-groupes résultant de la découpe d'un nœud \mathbf{t} en deux nœuds \mathbf{t}_L et \mathbf{t}_R . La variance d'un nœud \mathbf{t} étant définie :

$$V(\mathbf{t}) = \frac{1}{|\mathbf{t}|} * \sum_{(i \in \mathbf{t})} (y_i - \bar{y}_t)^2$$

Où,

$|\mathbf{t}|$: L'effectif du nœud \mathbf{t}

\bar{y}_t : Valeur moyenne de la variable \mathbf{Y} dans le nœud \mathbf{t}

Ainsi, en parcourant toutes les coupures (\mathbf{i}, \mathbf{d}) , on garde la coupure qui minimise la variance intra-groupe :

$$\frac{|t_L|}{n} * V(t_L) + \frac{|t_R|}{n} * V(t_R)$$

Dans le cas de la classification (l'ensemble des classes étant $\{1, \dots, L\}$). On définit l'impureté d'un nœud fils, le plus souvent par le biais de l'indice de Gini. L'indice de Gini d'un nœud t est défini :

$$\Phi(t) = \sum_{c=1}^L \widehat{p}_t^c * (1 - \widehat{p}_t^c)$$

Où,

\widehat{p}_t^c : La proportion d'observations de classe c dans le nœud t

On cherche alors toutes le couple (i, d) maximisant la pureté de l'indice de Gini définie par :

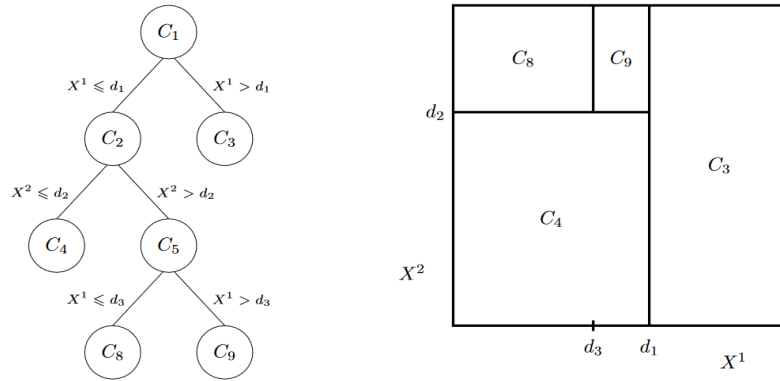
$$\Phi(t) - \left(\frac{|t_L|}{n} * \Phi(t_L) + \frac{|t_R|}{n} * \Phi(t_R) \right)$$

En classification comme en régression, l'idée principale étant d'augmenter l'homogénéité des nœuds obtenus. Notant que dans le cas d'une variable explicative X_i catégorielle, rien de ce qui précède ne change, dans ce cas, une coupure est simplement un élément de la forme :

$$\{X_i \in d\} \cup \{X_i \in \bar{d}\}$$

Où, d et \bar{d} sont des ensembles non vides qui constituent une partition de l'ensemble des modalités de la variable X_i .

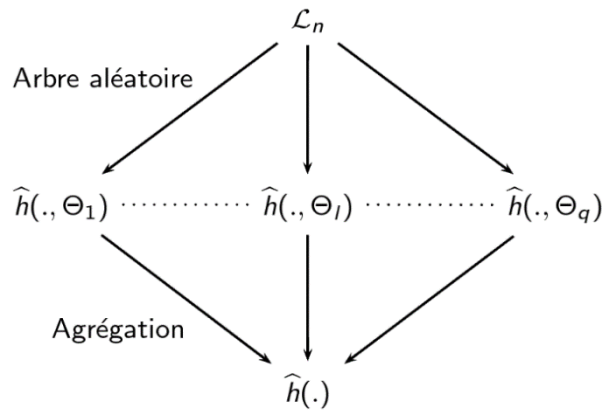
Une fois la racine de l'arbre est découpée, on se restreint à chacun des nœuds fils et on recherche alors, suivant le même procédé, la meilleure façon de les découper en deux nouveaux nœuds, et ainsi de suite. Les arbres sont ainsi développés, jusqu'à atteindre une condition d'arrêt. Une règle d'arrêt classique consiste à ne pas découper des nœuds qui contiennent moins d'un certain nombre d'observations. Les nœuds terminaux, qui ne sont plus découpés, sont appelés les feuilles de l'arbre.



Définition générale des forêts aléatoires (Breiman 2001) :

Soit $(\hat{h}(\cdot, \Theta_1), \dots, \hat{h}(\cdot, \Theta_q))$ une collection de prédicteurs par arbres, avec $(\Theta_1, \dots, \Theta_q)$ q variables aléatoires i.i.d. indépendantes de L_n . Le prédicteur des forêts aléatoires \hat{h}_{RF} est obtenu en agrégeant cette collection d'arbres aléatoires de la façon suivante :

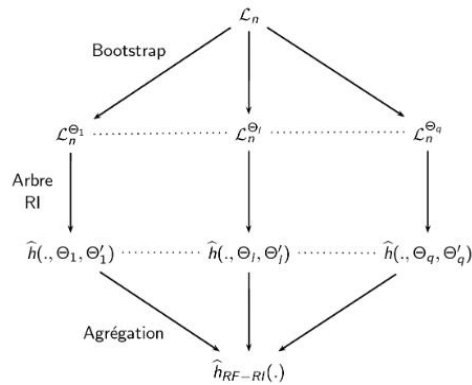
- $\hat{h}_{RF}(x) = \frac{1}{q} * \sum_{l=1}^q \hat{h}(x, \Theta_l)$ (moyenne des prédictions individuelles des arbres), en cas de régression.
- $\hat{h}_{RF}(x) = \operatorname{argmax}_{1 \leq K \leq L} \sum_{l=1}^q 1_{\hat{h}(x)=K}$ (vote majoritaire parmi les prédictions individuelles des arbres) en cas de classification.



Random Forests-RI : cas particulier des forêts aléatoires

Random Forests-RI signifie « forêts aléatoires à variables d'entrée aléatoires » (Random Forests with Random Inputs). Le principe de leur construction est tout d'abord générer plusieurs échantillons bootstrap (tirage avec remise) $L_n^{\Theta_1}, \dots, L_n^{\Theta_q}$. Ensuite, sur chaque échantillon $L_n^{\Theta_1}$ une variante de CART est appliquée. Dans ce cas pour découper un nœud,

on tire aléatoirement un nombre m de variables, et on cherche la meilleure coupure uniquement suivant les variables sélectionnées. La collection d'arbres obtenus est enfin agrégée (moyenne en régression, vote majoritaire en classification) pour donner le prédicteur Random Forests-RI.

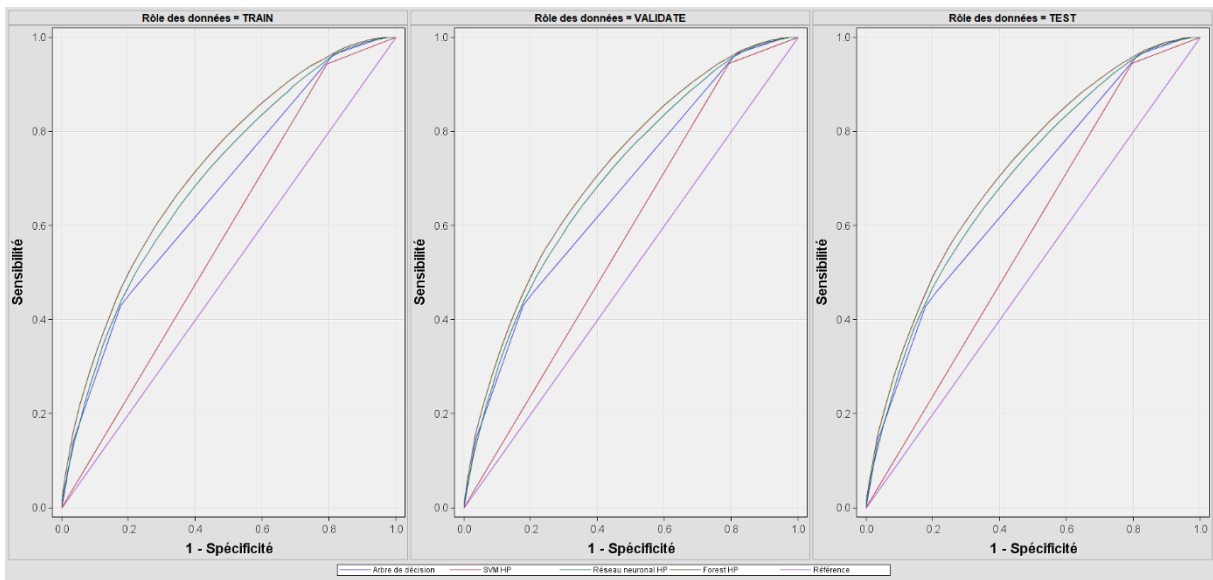


Sorties SAS :

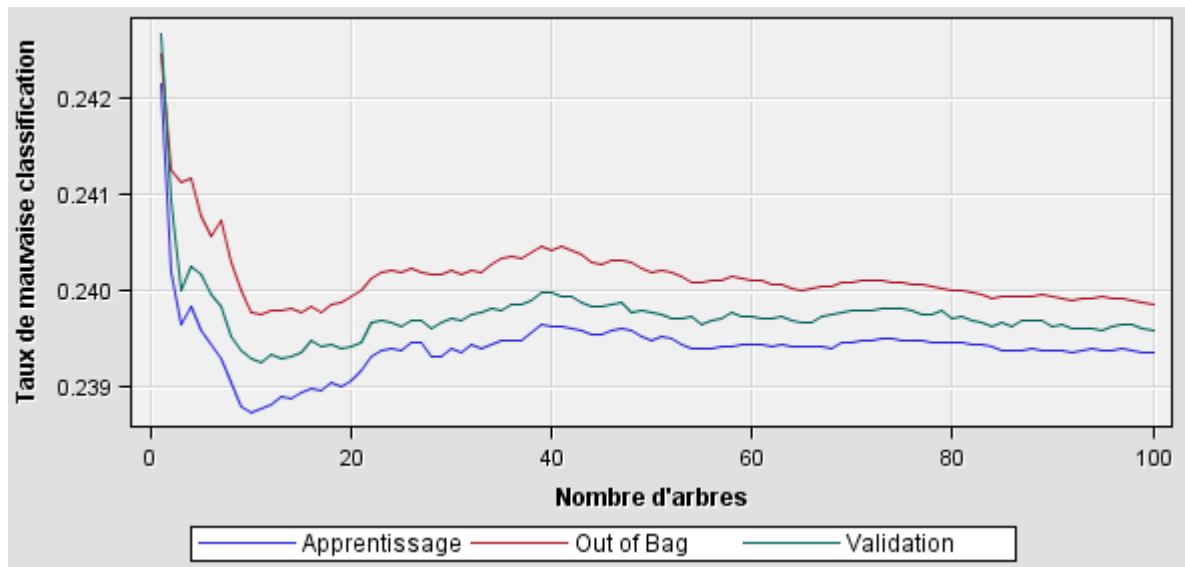
Importance des variables dans l'explication de Y :

Nom de la variable	Nombre de règles de découpe	Apprentis sage : Réduction de Gini	Apprentis sage : Réduction de marge	OOB : Réduction de Gini	OOB : Réduction de marge
prime_totale	12862	0.005448	0.010897	0.01671	0.01728
Anciennete_contrat	8524	0.008993	0.017986	0.04230	0.05628
Age_vehicule	7093	0.000980	0.001960	-0.00010	-0.00009
Age_conducteur	6516	0.001918	0.003837	0.02798	0.03035
Energie	5838	0.000549	0.001099	0.00973	0.01745
Puissance_fiscale	5336	0.000443	0.000885	0.00099	0.00049
Region	4697	0.000527	0.001054	-0.04602	-0.03429
DUSITF	4484	0.000460	0.000919	0.00421	0.00236
taux_CRM	4333	0.007709	0.015418	0.02563	0.03309
Type_echeance	3785	0.001463	0.002925	0.02978	0.02729
age_permis	3391	0.000580	0.001160	0.01022	0.00642
dist_BP	2664	0.000201	0.000402	-0.00137	-0.00261
Taux_chomage	2509	0.000135	0.000270	-0.02407	-0.02680
Type_garantie	2399	0.000134	0.000268	-0.00062	-0.00091
Taux_intermediaires	2232	0.000148	0.000296	-0.00943	-0.00890
Taux_activite	2224	0.000114	0.000229	-0.00984	-0.01108
nbr_veh_i_par_men...	2196	0.000132	0.000264	-0.03401	-0.03732
densite_veh_i	1912	0.000129	0.000259	-0.01861	-0.01619
sexe	1595	0.000079	0.000158	0.00098	0.00136

Courbes ROC :



Evolution du taux d'erreur en fonction du nombre d'arbres (Random Forest-RI) :



Annexe II : Etude d'existence d'une solution du programme(P) sous la contrainte
(3)

On rappelle que le programme (P) s'écrit :

$$(P): \begin{cases} \max_{P_1, P_2, \dots, P_n} \sum_{i=1}^n (P_i - AC_i) * f(P_i, \mathbf{X}_i) \\ P_i \leq \overline{P}_i \quad \forall i \in \{1, 2, \dots, n\} \quad (1) \\ P_i \geq \underline{P}_i \quad \forall i \in \{1, 2, \dots, n\} \quad (2) \\ \frac{1}{n} \sum_{i=1}^n f(P_i, \mathbf{X}_i) \geq \gamma \quad (3) \end{cases}$$

« Il existe toujours une solution du programme (P) sous la contrainte globale (3) »

Le lagrangien du programme (P) sous la contrainte (3) s'écrit :

$$\mathcal{L}(P_1, P_2, \dots, P_n, \lambda) = \sum_{i=1}^n (P_i - AC_i) * f(P_i, \mathbf{X}_i) - \lambda \left(\frac{1}{n} \sum_{i=1}^n f(P_i, \mathbf{X}_i) - \gamma \right)$$

Où λ est le multiplicateur de Lagrange associé à la contrainte (3).

A l'optimum on a :

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial P_i} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \end{cases} \Rightarrow \begin{cases} f(P_i^*, \mathbf{X}_i) + (P_i^* - AC_i) \frac{\partial f}{\partial P_i}(P_i^*, \mathbf{X}_i) - \frac{\lambda^*}{n} \frac{\partial f}{\partial P_i}(P_i^*, \mathbf{X}_i) = 0 \quad \forall i \in \{1, 2, \dots, n\} \quad (*) \\ \frac{1}{n} \sum_{i=1}^n f(P_i^*, \mathbf{X}_i) = \gamma \end{cases}$$

On obtient par la suite un système non linéaire de $n + 1$ équations avec $n + 1$ inconnues.

On sait que : $\forall i \in \{1, 2, \dots, n\} \quad \frac{\partial f}{\partial P_i} < 0$ on déduit alors que $\lambda^* > 0$.

On peut réécrire l'équation (*) comme suit :

$$\forall i \in \{1, 2, \dots, n\} \quad P_i^* = \frac{e_i(P_i^*)}{e_i(P_i^*) - 1} (AC_i + \delta^*) \quad \text{avec} \quad \delta^* = \frac{\lambda^*}{n} \quad (**)$$

On remarque que : $\forall i \in \{1, 2, \dots, n\} \quad e_i(P_i^*) > 1 \Leftrightarrow P_i^* \in \left[\left(\frac{-1}{\alpha_i(1+\beta_1)} \right)^{-1/\beta_1}; +\infty \right[$

Ainsi les primes sont bornées inférieurement.

Soit la fonction suivante :

$$g(P_i) = P_i - \frac{e_i(P_i)}{e_i(P_i) - 1} (AC_i + \delta^*)$$

L'équation (**) admet une solution sur $I = \left[\left(\frac{-1}{\alpha_i(1+\beta_1)} \right)^{-1/\beta_1}; +\infty \right[$ si la fonction g s'annule sur cet intervalle, étudiant alors la variation de cette fonction sur I :

$$\begin{cases} \frac{dg}{dP_i} = 1 + \frac{\partial e_i}{\partial P_i} \frac{AC_i + \delta^*}{(e_i(P_i) - 1)^2} \\ \frac{\partial e_i}{\partial P_i} = \alpha_i \beta_1^2 P_i^{-(\beta_1+1)} (f(P_i, \mathbf{X}_i))^2 > 0 \end{cases}$$

On déduit que la fonction g est croissante sur I , on a de plus :

$$\lim_{P_i \rightarrow \left(\frac{-1}{\alpha_i(1+\beta_1)} \right)^{-1/\beta_1}} g(P_i) = -\infty$$

$$\lim_{P_i \rightarrow +\infty} g(P_i) = +\infty$$

D'après le théorème des valeurs intermédiaires la fonction g s'annule sur I , ainsi l'existence d'un optimum de \mathcal{L} . On peut montrer qu'il s'agit d'un maximum en vérifiant que le déterminant de la matrice hessienne du lagrangien au point $(P_1^*, P_2^*, \dots, P_n^*, \lambda^*)$ est positif.