Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : El Ahmer Oumaima

Titre : Mise en place des modèles prédictifs du renouvellement en assurance automobile

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

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : SAHAM Assurance

Signature :

Directeur de mémoire en entreprise :

Nom : Lalla Mariam Jamali

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Secrétariat

Signature du candidat

Bibliothèque :

Remerciements

Je tiens à remercier Madame Lalla Mariam Jamali, ma tutrice de stage, pour ses conseils et sa disponibilité à mon égard.

Je remercie mon encadrant, M. Frastaoui Mohamed, actuaire chez SAHAM ASSURANCE, à qui j'exprime ma reconnaissance pour l'encadrement qu'il m'a prodigué tout au long de ma période de stage.

Je remercie Monsieur Abderrahim Oulidi, Directeur de l'Ecole d'Actuariat de Rabat, pour ses conseils pendant mon cursus académique et son encadrement du mémoire.

Je tiens à remercier Monsieur Yahia Salhi, professeur et enseignant chercheur à l'ISFA, pour m'avoir orienté dans mes recherches, pour ses précieux conseils et pour sa disponibilité.

En fin, je tiens à remercier ma camarade de l'UIR, Amina Ridouan pour l'aide apportée.

Mots clés : assurance automobile, analyse discriminante, score d'attrition, arbre de décision, modèle logistique, échantillonnage, *forward, ROC*, matrice de confusion.

Résumé

Dans un contexte économique de plus en plus exigeant en termes de performances financières, le risque lié au comportement de l'assuré prend toute son ampleur dans l'atteinte des objectifs de chaque compagnie.

Ce mémoire a été réalisé sur la base d'un contrat d'assurance automobile et a pour objet l'étude du comportement de renouvellement ou non du contrat d'assurance automobile.

Ainsi, après une étude préliminaire de la base de données, l'élaboration du modèle prédictif se fera à l'aide d'une analyse discriminante, une régression logistique et un arbre de décision. Chaque modèle sera appliqué sur un échantillon de travail extrait de la base de données, puis ce même modèle sera appliqué à un échantillon test afin de juger de sa pertinence.

Quant à l'outil technologique de l'étude, nous avons profité des possibilités fonctionnelles et techniques offertes par les plateformes *SAS 9.1*, *SAS Entreprise Guide* et *IBM SPSS*. Nous nous sommes appuyés sur des méthodes classiques de statistiques et modélisations actuarielles probabilisées et des techniques de *Data Mining*.

L'objectif de cette étude est d'apporter au Pôle d'Actuariat et Réassurance de la compagnie SAHAM Assurance une solution pratique pour une fidélisation de ses clients.

Key words: car insurance, discriminant analysis, attrition score, decision tree, logistic model, sampling, forward, ROC, confusion matrix.

Abstract

In an economic context that is increasingly demanding in terms of financial performance, the risk associated with the insured's behavior becomes more important in achieving the objectives of each company.

This study was carried out on the basis of an automobile insurance contract and aims to study the behavior of renewal or not of the contract of automobile insurance.

Thus, after a preliminary study of the database, the predictive model will be developed using discriminant analysis, logistic regression and decision tree. Each model will be applied to a work sample extracted from the database, and then the same model will be applied to a test sample to judge its relevance.

As for the study's technological tool, we took advantage of the functional and technical capabilities offered by the SAS 9.1 platform, SAS Enterprise Guide and IBM SPSS. We have relied on conventional methods of statistics and probabilistic actuarial modeling and Data Mining techniques.

The objective of this study is to provide SAHAM Assurance's Actuarial and Reinsurance Cluster with a practical loyalty solution for the client.

Note de Synthèse

La naissance de cette étude au sein du pôle actuariat et Réassurance de SAHAM assurance a pour but d'assurer une rentabilité du produit d'assurance automobile.
Elle s'inscrit dans l'optique de passer d'un marketing orienté produit à un marketing orienté client.

Notre travail repose sur deux axes principaux : la modélisation et la prévision.
Nous nous sommes appuyés sur les modèles prédictifs de datamining qui sont l'analyse discriminante, le modèle logistique et l'arbre de décision.

En raison des limites de certains modèles, nous avons pensé à combiner quelques modèles pour améliorer la performance du modèle, en occurrence, obtenir la plus faible probabilité de faire une mauvaise prédiction.

Nous souhaitons obtenir un modèle qui prédit le mieux l'acte de renouvellement d'un contrat d'assurance automobile, cependant, le taux d'erreur n'est pas toujours le meilleur moyen pour juger la pertinence d'un modèle. C'est la raison pour laquelle, nous avons opté pour d'autres mesures de fiabilité d'un modèle.

Pour mener à bien cette étude, la méthodologie utilisée est la suivante :

En premier lieu, nous appliquons chaque modèle avec l'ensemble de variables explicatives, par la suite nous testons le modèle avec un nombre réduit de variables. Ces variables sont choisies à partir de l'analyse discriminante comme outil de sélection de variables et le modèle logistique qui classe les variables les plus significatives.

En deuxième lieu, nous discrétisons les variables continues. Cette étape est importante pour appliquer le modèle logistique qui risque de ne pas bien tourner si les variables explicatives prennent plusieurs valeurs.

En fin, nous construisons nos modèles. Le 1er modèle est un modèle combinant les 3 modèles prédictifs, le 2ème modèle est le modèle arbre de décision appliqué sur le logiciel SPSS et le 3ème modèle est obtenu à partir de l'outil SAS Entreprise Guide.

Le tableau suivant montre les résultats pour chaque modèle utilisé.

Modèles	Taux d'erreur	
	Base d'apprentissage	Base test
Modèle 1	34.6%	33.6%
Modèle 2	30.7%	31.11%
Modèle 3	28.67%	31.19%

Le modèle 3 est le modèle ou le taux de faire une mauvaise prédiction est le plus bas sur la base d'apprentissage, cependant l'écart est important du taux d'erreur entre la base d'apprentissage et la base test.
Par ailleurs, le modèle le plus stable sur les deux bases est le modèle 2 qui est l'arbre de décision. Un modèle qui a sélectionné 4 variables les plus discriminantes qui sont CRM, profil de risque, prime et puissance fiscale.

Un modèle n'est pas pour autant unique, la performance d'un modèle dépend plus de la qualité des données et du type de problème que de la méthode.

Table des matières

Remerciements 3
Résumé 4
Abstract 5
Note de Synthèse 6
Avant-propos 12
I. Introduction 13
1. Présentation de l'organisme d'accueil : Saham Assurance Maroc 13
2. Historique 15
3. Pôle Actuariat et Réassurance 17
4. Problématique 18
II. Etude de la base de données 19
1. Vue globale de la base de données 19
2. Présentation des différentes variables 19
3. Analyse statistique 20
3.1 Création de nouvelles variables 20
3.2 Explorer la distribution des variables 21
3.3 Coefficient de corrélation 28
III. Modèles prédictifs pour l'attrition 32
1. Cadre théorique 32
1.1 Principe du Scoring 32
1.2 Limites du champ d'application du modèle classique 33
1.3 Modèles prédictifs 33
1.3.1 Analyse discriminante 34
1.3.2 Régression logistique 35
1.3.3 Arbre de décision 37
1.4 Comparaison des modèles 41
1.4.1 Matrice de confusion 41
1.4.2 Courbe ROC 43
2. Sélection des variables 44
2.1 Différentes méthodes 44
2.2 La méthode utilisée 45
3. Réalisation d'un échantillonnage 47
3.1 Pourquoi avoir recours à un échantillonnage ? 47
3.2 L'échantillonnage adopté pour l'étude 48
4. Application : Modélisation 51

4.1 Classification de variables continues 51
4.2. Analyse discriminante 53
4.3. Modèle 1 : Régression logistique, analyse discriminante et arbre de décision 55
4.4. Modèle 2 : Arbre de décision 59
4.5. Application sur la base test 61
Modèle 1 61
Modèle 2 62
4.6. Modèle 3 : Application sur SAS Entreprise Guide 63
Application 66
III. Conclusion 71
Bibliographie 72
Annexes 73
Annexe 1 : Variables spécifiques du contrat 73
Annexe 2 : Les données « Garanties » 75
Annexe 3 : Analyse statistique de quelques variables quantitatives 77
Annexe 4 : Arbre de décision « Discrétisation par l'algorithme CHAID » 78
Annexe 5: Résultats de la modélisation 79
- L'analyse discriminante 79
- Modèle logistique 80

Listes des Figures

Figure 1 : Données du GROUPE SAHAM 16
Figure 2: Représentation de la variable « renouvele » 22
Figure 3: Distribution de l'âge du conducteur pour « renouvele=Non » 23
Figure 4: Distribution de l'âge du conducteur pour "renouvele=Oui" 24
Figure 5: Distribution de la variable "Profil_Risque" 25
Figure 6: Distribution de la variable "CRM" 26
Figure 7: Les régions du portefeuille 27
Figure 8: Fréquence des assurés dans la région du « SAHARA » 27
Figure 9: Fréquence des assurés dans la région « Gharb-Chrarda-Beni Hssen » 28
Figure 10: Arbre de décision 39
Figure 11: Représentation de la courbe ROC 43
Figure 12: Récapitulatif des choix des variables selon chaque méthode de sélection 45
Figure 13: Résultats de la sélection "Forward" 46
Figure 14: Macro utilisée sur SAS pour la méthode d'échantillonnage "cross-validation" 49
Figure 15: Représentation du pourcentage de la variable « renouvele » sur la base d'apprentissage et base test 49
Figure 16: Discrétisation de la variable "CRM" par CHAID 52
Figure 17: Discrétisation de la variable "Profil_Risque" par CHAID 52
Figure 18: Matrice de confusion de l'analyse discriminante 54
Figure 19: Statistique F du modèle analyse discriminante 54
Figure 20: Variables non retenues de l'analyse discriminante 55
Figure 21: Impact global des variables explicatives du modèle « régression logistique » 56
Figure 22: Extrait du tableau des estimations des paramètres 57
Figure 23: Les 22 premières lignes du tableau des estimations des rapports de cotes 58
Figure 24: Association des probabilités prédites et valeurs estimées 58
Figure 25: Représentation du modèle arbre de décision 60
Figure 26: Résultats de classification de l'arbre de décision sur la base d'apprentissage 61
Figure 27: Coût d'un arbre de décision 61
Figure 28: Association des probabilités prédites et observées 62
Figure 29: Résultats de classification de l'arbre de décision sur la base test 62
Figure 30: Classement des variables sélectionnées 67
Figure 31: Matrice de confusion du modèle 3 68
Figure 32: Courbe ROC du modèle 3 68
Figure 33: Résultats des modèles sur SAS Rapid Predictiv Model 69
Figure 34: Statistiques du modèle 3 69
Figure 35: Discrétisation de la variable "age_conducteur" par arbre de décision 78
Figure 36: Discrétisation de la variable "prime" par arbre de décision 78
Figure 37: rapport de cote (odds ratio) de la régression logistique 81

Listes des tables

Table 1: Evolution du chiffre d'affaires par branche d'activité en MDhs 14
Table 2 : Evolution du chiffre d'affaires du marché d'assurance non vie par compagnie en MDhs 15
Table 3: Quelques Variables explicatives classées dans chaque catégorie 20
Table 4: Statistiques descriptives de la variable "age_conducteur" 24
Table 5: Statistique de la variable" Profil_Risque" 25
Table 6: Coefficients de corrélation de Spearman 29
Table 7: Corrélation croisées entre les variables 30
Table 8: Variables liées au contrat 73
Table 9: Les données liées au « conducteur » 74
Table 10: Liste des données garanties 76
Table 11: Analyse statistique de 4 variables quantitatives 77

Avant-propos

Ce monde concurrentiel nous incite à faire de notre mieux dans tous les aspects de notre existence, que ce soit nos études, notre travail et même notre niveau de vie. Ne se limitant pas au niveau individuel mais atteignant aussi le niveau collectif et entrepreneurial, c'est cette concurrence même qui pousse les différentes entreprises et compagnies à se différencier afin de présenter aux clients les meilleures offres.
Toute entreprise a pour but principal d'acquérir de nouveaux clients, ce qui explique la concurrence de plus en plus soutenue.

Les entreprises d'assurances cherchent à gagner en notoriété sur le marché mais aussi en clientèle, qui est l'essence de la continuité et de la réussite de toute firme. Elles se trouvent dans un marché où sont aussi présentes les sociétés d'assurance mutuelle, les institutions de prévoyance, les bancassurances et sont donc confrontées à bien connaître leurs clients pour concevoir une stratégie de fidélisation basée et bien étudiée sur leurs comportements afin de prétendre à leur survie.

En effet, le coût d'acquisition d'un nouveau client est très élevé : Il a été estimé qu'en moyenne, il coûte 100$ pour acquérir un nouveau client tandis qu'il coûte que 30$ pour le conserver. De plus, une augmentation de la rétention de l'ordre de 5% amène une augmentation de 80% de la valeur à vie d'un client.

Aujourd'hui, toute assurance se retrouve face à la question suivante : dans mon portefeuille présent, quels sont les clients qui renouvelleront leurs contrats et quels sont ceux qui risquent de ne pas le faire ?

I. Introduction

1. Présentation de l'organisme d'accueil : Saham Assurance Maroc

L'assurance au Maroc est dominée avec chacune 30% du chiffre d'affaires du secteur par l'automobile et l'assurance vie. Le secteur est considéré comme étant l'un des secteurs les plus dynamiques et les plus matures de l'économie nationale en vue de sa croissance perpétuelle, se positionnant ainsi à la deuxième place dans le continent africain et à la troisième dans la région Moyen-Orient et Afrique du nord (MENA[1]) en termes de chiffres d'affaires.

Le marché de l'assurance marocaine est un marché ouvert, diversifié et fortement concurrentiel comptant 18 compagnies d'assurance dont Wafa Assurance, RMA Watanya, Axa Assurance Maroc et Saham Assurance sont les leaders. Et selon les données compilées et récentes de la Fédération Marocaine des Sociétés d'Assurance et de Réassurance (FMSAR) relatives aux réalisations du secteur au terme des 6 premiers mois de 2014, les primes émises se sont établies à 15,79 milliards de dirhams.

Par branche d'activité, c'est l'Automobile qui domine avec des primes de l'ordre de 5,20 milliards de dirhams, soit une part de marché de 32,9 % devant l'Assurance Vie & capitalisation (29,1 %), les Accidents corporels (10,80 %), les Accidents de travail (8,90 %) et les Incendies (5,8 %). Ces cinq branches représentent environ 88 % des primes du secteur. Par activité, la non-Vie demeure prépondérante avec un total de 11,07 milliards de dirhams de primes émises au titre du premier semestre 2014, soit 71 % des primes globales du secteur. Elle est largement dominée par la branche «Automobile» avec des primes ressortant à 5,20 milliards de dirhams, soit 47 % du total de l'activité non-Vie, avec une intense concurrence entre tous les acteurs du marché. SAHAM Assurance domine ce segment avec une part de marché de 19,6 %, suivie par Wafa Assurance (16,9 %), Axa Assurance Maroc (14,03 %) et RMA Watanya (13,5 %).

Les tableaux suivants montrent la structure du chiffre d'affaire du secteur de l'assurance au Maroc :

[1] MENA: Acronyme de "Middle East and North Africa"

	2013	2014	2015	Evolution 2014/2015	Evolution 2013/2014
Assurances Vie & Capitalisation	**8 598,6**	**9 399,1**	**10 560,8**	**12,4%**	**9,3%**
Assurances Individuelles	5 192,3	5 641,3	6 308,5	11,8%	8,6%
Assurances de Groupes	1 987,8	2 061,4	2 106,5	2,2%	3,7%
Capitalisation	1 257,7	1 368,2	1 684,3	23,1%	8,8%
Contrats à Capital Variable	152,7	326,7	460,1	40,9%	114%
Acceptations Vie	8,1	1,4	1,4	-5,8%	-82,2%
Assurances Non Vie	**18 135,0**	**19 022,5**	**19 862,9**	**4,4%**	**4,9%**
Accidents Corporels	3 068,8	3 224,0	3 359,5	4,2%	5,1%
Accidents du Travail	2 140,4	2 213,5	2 090,9	-5,5%	3,4%
Automobile	8 497,1	9 033,7	9 514,2	5,3%	6,3%
Responsabilité Civile Générale	509,5	509,3	544,4	6,9%	0,0%
Incendie	1 255,4	1 159,3	1 312,1	13,2%	-7,7%
Risques Techniques	377,1	416,0	393,7	-5,4%	10,3%
Transport	587,4	568,5	552,3	-2,9%	-3,2%
Autres Opérations Non Vie	606,7	606,3	701,2	15,7%	-0,1%
Assistance - Crédit - Caution	968,9	1 091,1	1 183,2	8,4%	12,6%
Acceptations Non Vie	123,8	200,9	211,5	5,3%	62,2%
Total	**26 733,6**	**28 421,6**	**30 423,7**	**7,0%**	**6,3%**

Source : FMSAR

Table 1: Evolution du chiffre d'affaires par branche d'activité en MDhs

	2013	2014	2015	Evolution 2014/2015	Part marché
Atlanta	1 271,4	1 410,8	1 550,7	9,9%	7,8%
Axa Assistance Maroc	88,4	125,2	152,4	21,7%	0,8%
Axa Assurance Maroc	2 793,5	2 762,0	2 795,2	1,2%	14,1%
CAT	648,7	636,1	631,2	-0,8%	3,2%
Coface Maroc	-	-	17,9	-	-
Euler Hermes ACMAR	90,5	96,3	108,3	12,5%	0,5%
MAMDA	725,2	759,9	846,9	11,4%	4,3%
Maroc Assistance Internationale	391,4	436,5	432,6	-0,9%	2,2%
Marocaine Vie	62,7	68,5	83,0	21,1%	0,4%
MATU	221,3	247,0	265,1	7,3%	1,3%
MCMA	423,1	471,2	551,8	17,1%	2,8%

Mutuelle Taamine Chaabi	-	-	-	-	-
Rma Watanya	2 625,0	2 726,2	2 874,7	5,4%	14,5%
Saham Assistance	297,6	298,4	333,2	11,7%	1,7%
Saham Assurance	3 108,5	3 309,0	3 410,3	3,1%	17,2%
Sanad	1 282,7	1 337,8	1 420,0	6,2%	7,1%
Wafa Assurance	2 919,2	3 059,0	2 985,0	-2,4%	15,0%
Wafa IMA Assistance	106,6	133,8	176,7	32,0%	0,9%
Zurich Assurance Maroc	1 079,1	1 144,7	1 228,0	7,3%	6,2%
Total	**18 135,0**	**19 022,5**	**19 862,9**	**4,4%**	**100,0%**

Table 2 : Evolution du chiffre d'affaires du marché d'assurance non vie par compagnie en MDhs

Les primes émises au titre du premier semestre de 2015 par le secteur marocain des assurances ont atteint 16,8 Mds de DH, soit une progression de 6,3% par rapport à fin juin 2014. C'est ce que révèlent les statistiques de la Fédération marocaine des sociétés d'assurances et de réassurance.

La répartition du chiffre d'affaires montre que l'assurance non-vie s'accapare une part de 11,55 Mds de DH, soit +3,2%, tandis que l'assurance vie et capitalisation enregistre une hausse de 13,9% à 5,2 Mds de DH.
La compagnie Wafa Assurance occupe toujours la position de leader du secteur avec des primes émises (vie et non-vie) de 3,4 Mds de DH, en hausse de 6,5%, pour une part de marché de 20,4%. Elle est suivie par RMA Watanya avec 2,85 Mds de DH (+10,1%) et une PDM de 17%. Avec un chiffre d'affaires constant d'un semestre à l'autre (2,16 Mds de DH) et une part de marché de 12,9%, Axa Assurance Maroc occupe la troisième marche du podium, talonnée par Saham Assurance (+1,7% à 2,11 Mds de DH) qui détient une PDM de 12,6%.

2. Historique

Naissance de SAHAM ASSURANCE Maroc :

Groupe marocain fondé en 1995, SAHAM est un acteur de référence dans plusieurs secteurs d'activité et en particulier dans les métiers de service à forte valeur ajoutée notamment ceux de l'Assurance, l'*Offshoring*, la santé et l'immobilier.

Depuis sa création, SAHAM connait une croissance continue qui hisse le Groupe au rang de leader dans ses domaines de prédilection. En 2012, le Groupe a enregistré un chiffre d'affaire de 900 millions de dollars, soit 7753 milliards de dirhams, et emploie plus de 5900 personnes en Afrique et au Moyen-Orient.

Quelques dates clés :

2005 : Acquisition par le Groupe SAHAM de 67,01 % de CNIA Assurance

2007 : Acquisition par le Groupe SAHAM des Assurances ES-SAADA

2009 : Fusion entre CNIA Assurance et les Assurances ES-SAADA, formant ainsi CNIA SAADA Assurance.

2010 : Introduction en bourse de CNIA SAADA Assurance. Acquisition du Groupe COLINA. Le rachat de COLINA par le Groupe

Saham a permis au Groupe d'acquérir une plateforme d'expansion en Afrique Sub-Saharienne. Le Groupe COLINA est le premier assureur de la zone africaine CIMA (Conférence Interafricaine des Marchés d'Assurance).

2011 : Acquisition de l'unité de production de GalaxoSmithKline Maroc qui constitue un premier jalon du développement du pôle Santé du Groupe.

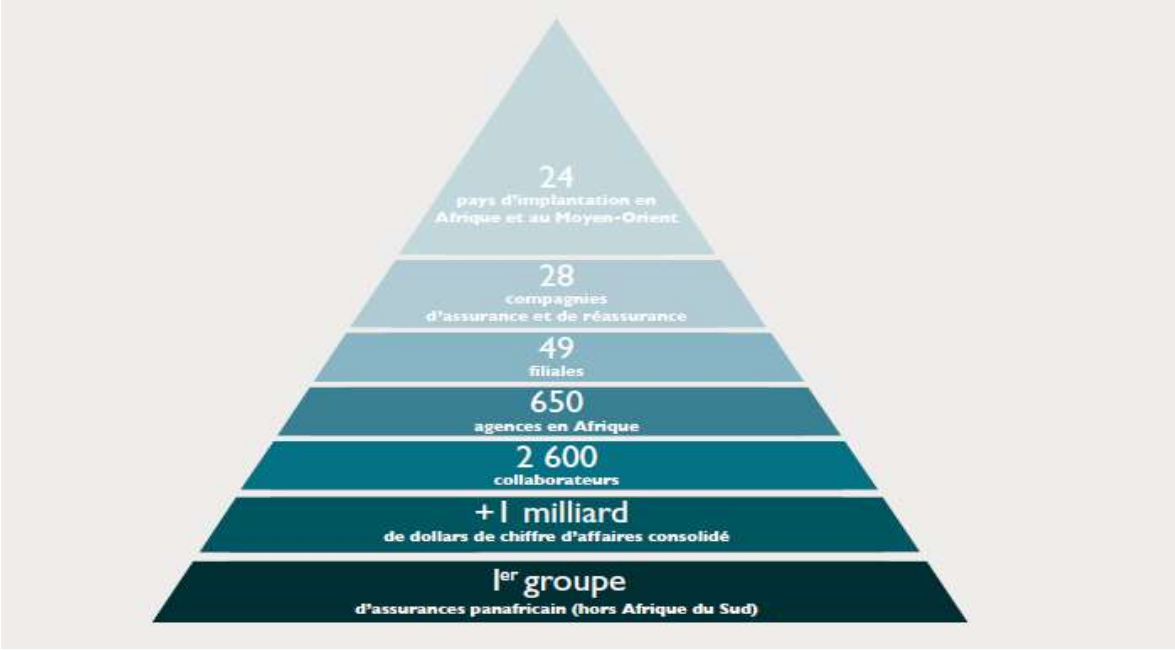


Figure 1 : Données du GROUPE SAHAM

3. Pôle Actuariat et Réassurance

L'entité Actuariat de ce pôle anticipe les risques financiers que prend la compagnie dans la souscription et la gestion des contrats et s'assure de sa solvabilité. L'entité Réassurance est garante de la protection de la compagnie par un bon transfert des risques. En 2012, ce pôle a mené à bien l'étude, l'analyse et la mise en place des pré-requis pour le passage aux normes IFRS (*International Financial Reporting Standards*).

Concrètement, au niveau du pôle, ce projet a nécessité la mise en place de tests actuariels de suffisance des provisions techniques. Avec les normes IFRS, l'harmonisation internationale de l'information financière dans tous ses aspects, y compris comptables, va faire ressortir la réalité économique de la compagnie et facilitera la comparabilité entre les sociétés. Concernant l'entité Réassurance, les traités ont été renouvelés et les travaux d'optimisation commencés l'année précédente au niveau de la structure, du coût et de la sécurité ont été poursuivis. Cela a permis d'améliorer le coût de la réassurance et de réduire la cession de primes de la compagnie.

En automobile, dans un marché en constante évolution, le pôle est resté très réactif dans les travaux d'ajustements des produits offerts, ce qui lui a permis de demeurer compétitif et rentable sur une branche qui constitue le cœur du métier de la compagnie.

4. Problématique

L'objectif de notre étude est d'établir une modélisation pour choisir le modèle qui prédit le mieux le renouvèlement du contrat d'assurance automobile. Par la suite cette modélisation contribuera à une discussion pour la stratégie interne de l'entreprise.

L'enjeu ici est donc de pouvoir estimer, en se basant sur les différentes informations détenues dans la base de donnée la probabilité de non renouvellement du contrat afin d'avoir une idée globale sur le portefeuille futur et pouvoir identifier les déterminants pouvant causer l'attrition de telle ou telle personne, ceci dans le but de réfléchir sur la politique à adopter pour lui faire changer d'avis et la garder dans le portefeuille.

Dans ce cadre, notre travail consistera à utiliser les méthodes de *data-mining*, pour estimer à partir de la base de données fournie par notre organisme encadrant, SAHAM Assurance, les probabilités de non renouvellement de sa clientèle et ce en adoptant des méthodes d'analyse prédictive comme l'arbre de décision, la régression logistique et l'analyse discriminante.

Nous nous sommes principalement basées sur les logiciels SAS base, SAS Entreprise Miner et IBM SPSS pour nos différents traitements.

- En premier lieu, nous décrivons l'importance des données extraites et les premières intuitions procurées par des analyses descriptives sur un portefeuille d'assurance automobile sur les années 2014 et 2015.

- En deuxième partie, nous décrirons la théorie du *scoring* « score d'attrition » ainsi que les modèles sur lesquels nous nous sommes basés pour prédire le renouvellement ou non d'un contrat d'assurance automobile.

- Par la suite, une classification des variables continues est appliquée en s'appuyant sur les arbres de décision. Nous évoquons les étapes préliminaires de la modélisation, que nous illustrerons avec l'application des modèles prédictifs sur SAS et IBM SPSS.

- La troisième partie est dédiée à la comparaison des modèles avec la matrice de confusion et courbe ROC et à l'application sur une base que nous appelons test pour juger la pertinence du modèle.

- En conclusion, une synthèse sur la pertinence des modèles obtenus sera présentée. Puis des voies d'amélioration du modèle seront proposées.

II. Etude de la base de données

La première partie du travail repose sur l'analyse primaire de la base automobile. A travers cette analyse, on a été amené à explorer la distribution des variables en vérifiant la fiabilité des variables : valeurs incohérentes ou manquantes, suppression ou isolement de certaines variables non suffisamment renseignées. Eventuellement, « effectuer une discrétisation supervisée » : découper la variable en tranches en fonction de la variable à expliquer et isoler les valeurs manquantes ou aberrantes.

Dans un second temps, les statistiques bi-variées. Cela nous a permis :

- D'analyser les incohérences entre variables ainsi que la liaison entre notre variable « cible » et les variables explicatives.
- S'assurer que la distribution des valeurs de chaque variable est bien homogène et ne contient pas de valeur erronée.

En fin, pour ne pas fausser les modèles qu'on appliquera il faut s'assurer que les variables entre elle ne sont pas très corrélées. Cela est important pour ne pas fausser les modèles.

1. Vue globale de la base de données

Notre base contient 1368133 de contrats d'assurance automobile sur l'année 2014 et 2015.
Elle contient 1368133 lignes, chaque ligne étant un client et 165 colonnes qui regroupent des données sur ce client.

La base automobile qui m'a été remise était sous forme texte «txt », comportant les informations sur l'exercice de l'année 2014 et 2015. Elle contient l'ensemble des informations sur chaque police (Date naissance, prime …). Il a donc été question de façonner un programme d'importation sous SAS en respectant le type de chaque variable pour ne pas fausser les informations.
La base finale doit obtenir l'ensemble des informations nécessaire pour modéliser et prédire le comportement de renouvellement.

2. Présentation des différentes variables

Dans le cadre de l'analyse, nous avons pu classer nos variables selon quatre catégories :

- Les données « Contrats »

Cette catégorie comprend des variables liées au conducteur et d'autres liées à l'agent ou l'intermédiaire. Elle contient aussi des variables spécifiques au contrat ainsi que des variables de gestion.

- Les données « Garanties »

Dans cette section, on a les variables ayant rapport avec les garanties fournies dans les différents contrats. Nous détaillerons les options, les différentes garanties et leurs primes.

- Les données « Véhicule »

Il s'agit ici de toute variable décrivant le véhicule et ses caractéristiques.

- Les données « Annexes »

Il s'agit de données que nous n'avons pas pu attribuer à aucune des autres catégories.

Le tableau ci-dessous représente quelques variables classées pour chaque type.

Contrats	Garanties	Véhicule	Annexes
Anciennete_permis	Actes_de_Vandalismes	Marque	EXERCICE
CRM	Collision	Model	FLOTTE_TYPE
CSP	Incendie	Poids_en_Charge	GROUPE
Date_Naissance	Inondation	Combustion	ANNEE_EFFET
Sexe	MRH	ANCIENNETE_MEC	IND_FLOTTE
Situation_matrimoniale	Vol	MODEL	
Prime	Bris_de_glaces	VALEUR_NEUVE	

Table 3: Quelques Variables explicatives classées dans chaque catégorie

Etant donné que l'efficacité d'un modèle se base sur la pertinence des variables et non leur quantité, nous essayerons d'en diminuer le nombre de façon à garder l'information transmise et donc assurer l'efficacité du modèle.
Se référer à l'annexe pour voir l'ensemble de variables retenues.

3. Analyse statistique

3.1 Création de nouvelles variables

Pour en tirer un apport concret sur la liaison entre la variable du renouvellement et les variables explicatives, on a créé de nouvelles variables que nous citerons ci-dessous :

- Renouvele_b (la variable à expliquer) : la variable binaire qui prend 1 si l'assuré a renouvelé son contrat ou 0 sinon.

- Mois_fin : la variable « fin» nous fournit des informations sur l'année fin de chaque contrat. Chaque observation ayant une date propre à elle, cela n'ajoutera pas de plus-value à notre étude. Nous avons donc vu qu'il était plus pertinent de retenir le mois dans lequel finit le contrat. Nous aurons ainsi 12 modalités pour cette variable et plus de justesse pour notre étude.

- gar_ann: Nous avons 30 variables exprimant les différentes garanties auxquelles tout contractant a le droit de souscrire. Il n'est évidemment pas raisonnable d'inclure ce grand nombre, donnant les détails sur les garanties dans notre étude. Nous avons vu qu'il était plus judicieux de créer une nouvelle variable que nous avons nommé « gar_ann» qui exprimera si oui ou non le contractant a souscrit à l'une des différentes garanties. Cette variable nous donnera donc une vue d'ensemble sur les garanties annexes sans trop en détailler le contenu.

- Nous créons trois autres variables qui sont « age_permis », « age_conducteur » et « age_vehicule » à partir des variables « DATE_EXIGIBILITE », « anciennete_permis », « Date_Naissance » et « ANCIENNETE_MEC ».

3.2 Explorer la distribution des variables

La variable cible est la variable à expliquer binaire (renouvle_b) qui a été créé de la variable existante dans la base de donnée « renouvele » qui prend les modalités oui ou non.

Elle prend 0 si l'assuré n'a pas renouvelé son contrat et 1 sinon.

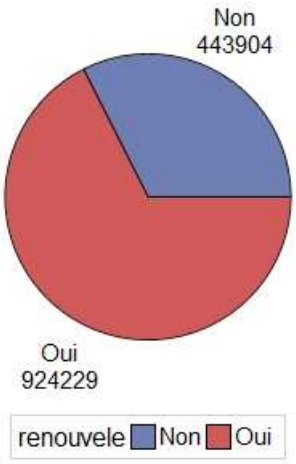


Figure 2: Représentation de la variable « renouvele »

32.45% d'assurés n'ont pas renouvelé leurs contrats d'assurance automobile sur les années 2014 et 2015.

Notre base comporte 61228 de femmes et 845189 d'hommes et 109439 non renseignés.

Sexe = F

renouvele	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
Non	18776	30.67	18776	30.67
Oui	42452	69.33	61228	100.00

Sexe= M

renouvele	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
Non	276492	32.71	276492	32.71
Oui	568697	67.29	845189	100.00

On constate que les hommes ont tendance plus à ne pas renouveler leurs contrats d'assurances automobile plus que les femmes.

- Variable « age_conducteur » :

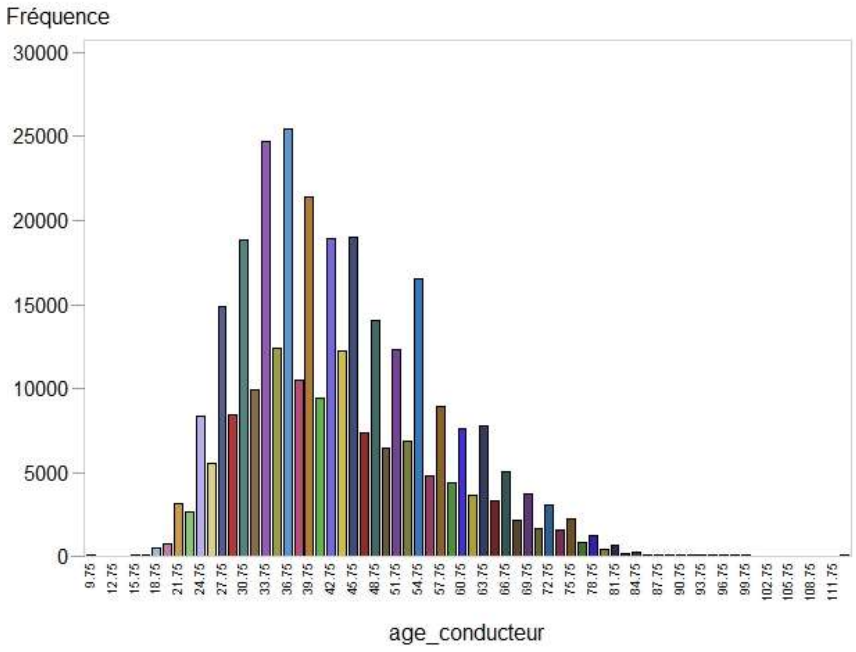


Figure 3: Distribution de l'âge du conducteur pour « renouvele=Non »

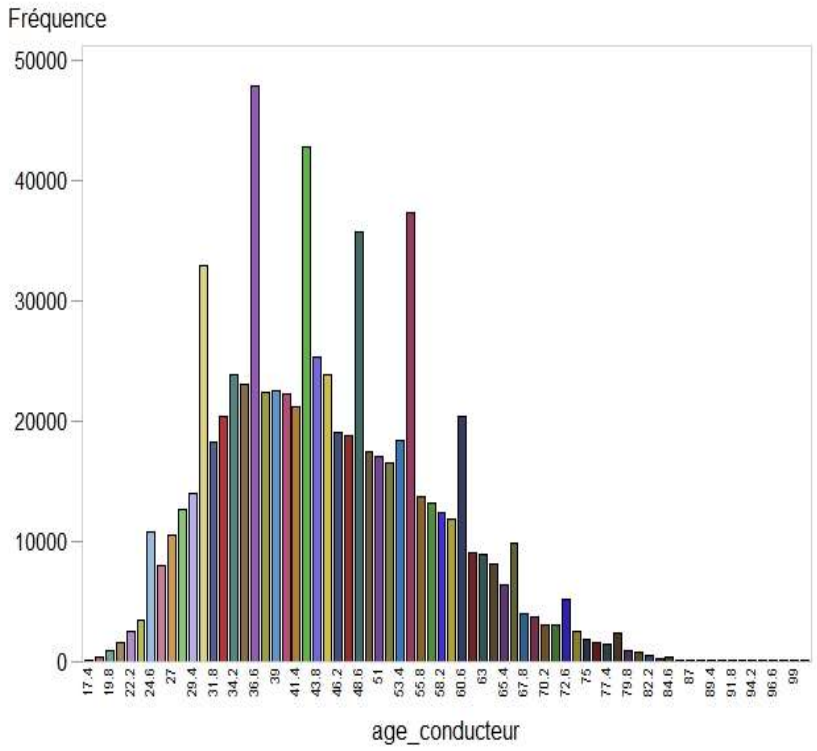


Figure 4: Distribution de l'âge du conducteur pour "renouvele=Oui"

Variable d'analyse : age_conducteur						
renouvele	N Obs	Moyenne	Ecart-type	Minimum	Maximum	N
Non	443904	43.5994967	12.9133621	10	113	443904
Oui	924229	44.9746134	12.1884435	17	100	924229

Table 4: Statistiques descriptives de la variable "age_conducteur"

L'âge moyen du conducteur de notre portefeuille est de 44.5 ans.

Variable profil de risque « profil_risque »

Saham Assurance a créé un indice pour la variable « profil-risque » afin d'améliorer la politique de fidélisation des clients.

Cette variable « Profil_Risque » prend les chiffres de 0 à 5.

1 correspond à un client excellent, 2 bon, 3 client moyen, 4 client risqué et 5 un client mauvais « trop risqué »

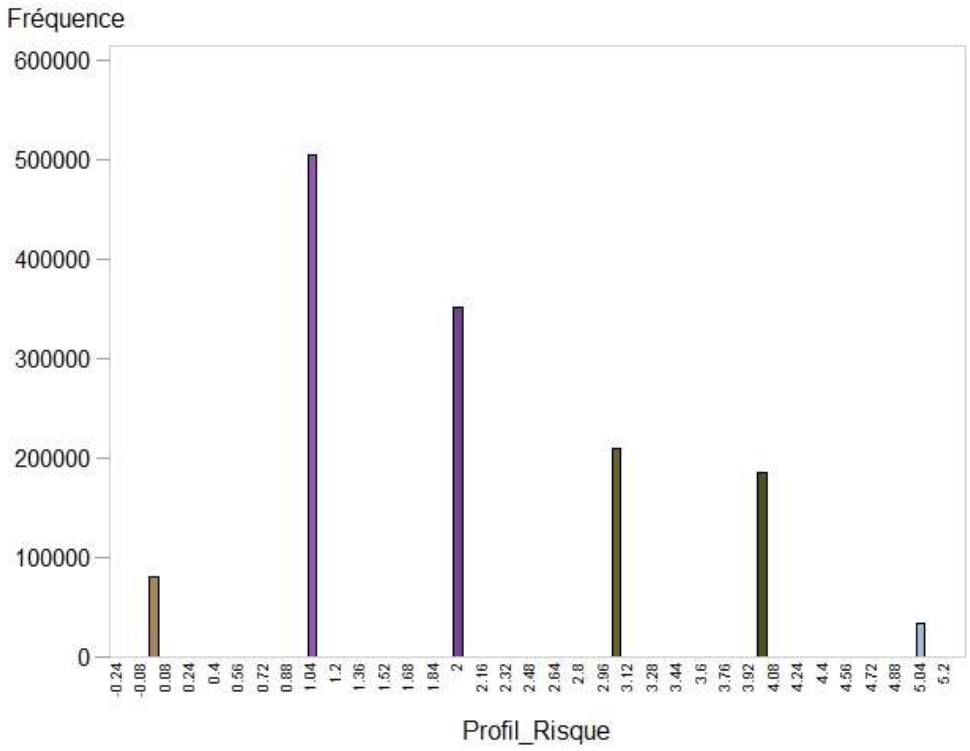


Figure 5: Distribution de la variable "Profil_Risque"

Variable d'analyse : Profil_Risque						
renouvele	N Obs	Moyenne	Ecart-type	Minimum	Maximum	N
Non	443904	2.1912328	1.2565438	0	5.0000000	443904
Oui	924229	1.9286724	1.2178527	0	5.0000000	924229

Table 5: Statistique de la variable" Profil_Risque"

Nous constatons que la majorité de clients est classée entre 1 et 3.

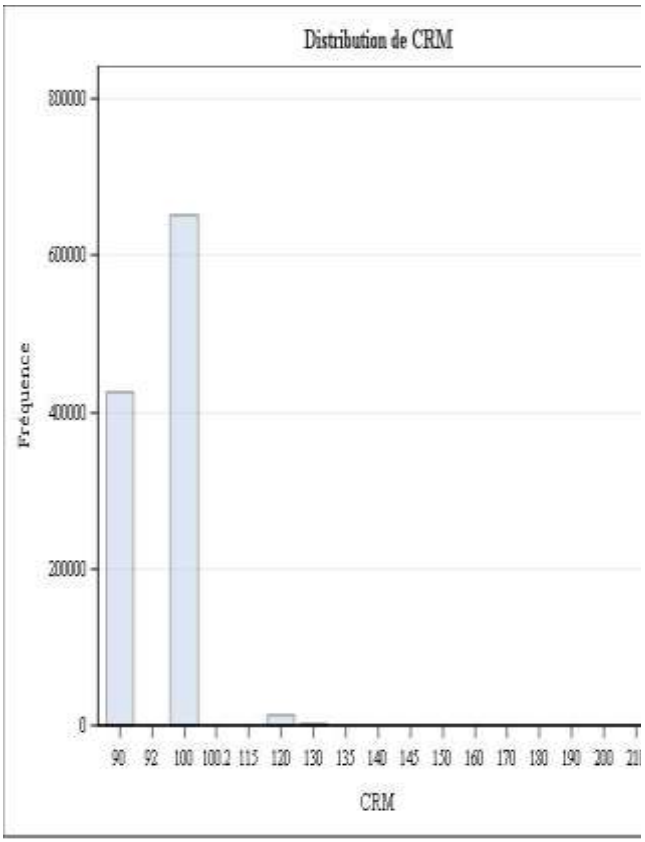
Variable « CRM »

Le CRM (coefficient de réduction majoration) au Maroc est apparu en 2006 remplaçant le système bonus-malus. Il agit sur la prime d'assurance. Selon la fédération marocaine d'assurance, ce coefficient est compris entre 90% et 250% et fonctionne comme suit :

- La prime est réduite de 10% si aucun accident n'a eu lieu durant une période de 24 mois consécutifs précédant la souscription ou le renouvellement du contrat.

- En cas d'accident engageant ou susceptibles d'engager totalement ou partiellement la responsabilité durant les 12 mois précédant la souscription ou le renouvellement du contrat, la prime est majorée de 20% pour chaque accident matériel et de 30% pour chaque accident corporel sans dépasser 250%.

Ces taux sont respectivement de 15% et de 20% pour l'exploitation d'un véhicule destiné au transport public de voyageurs ou un souscripteur d'une responsabilité civile garagiste.



CRM	Fréquence
90	425645
92	3
100	651499
100.2	1
115	14
120	13390
130	2879
135	1
140	670
145	1
150	96
160	69
170	4
180	7
190	1
200	1
210	1

Figure 6: Distribution de la variable "CRM"

On remarque que dans notre portefeuille, le CRM est compris entre 90% et 210%, la majorité des assurés ont un CRM de 100%, 32.14% d'assurés leur prime est minorée de 10% et 10.22% d'assurés leur prime est majorée de 1.2 %.

Variable « Region »

Le Maroc comptait 16 régions, cependant depuis 2015 le nombre de région est réduit à 12.

14 régions sont représentées sur la Figure 7

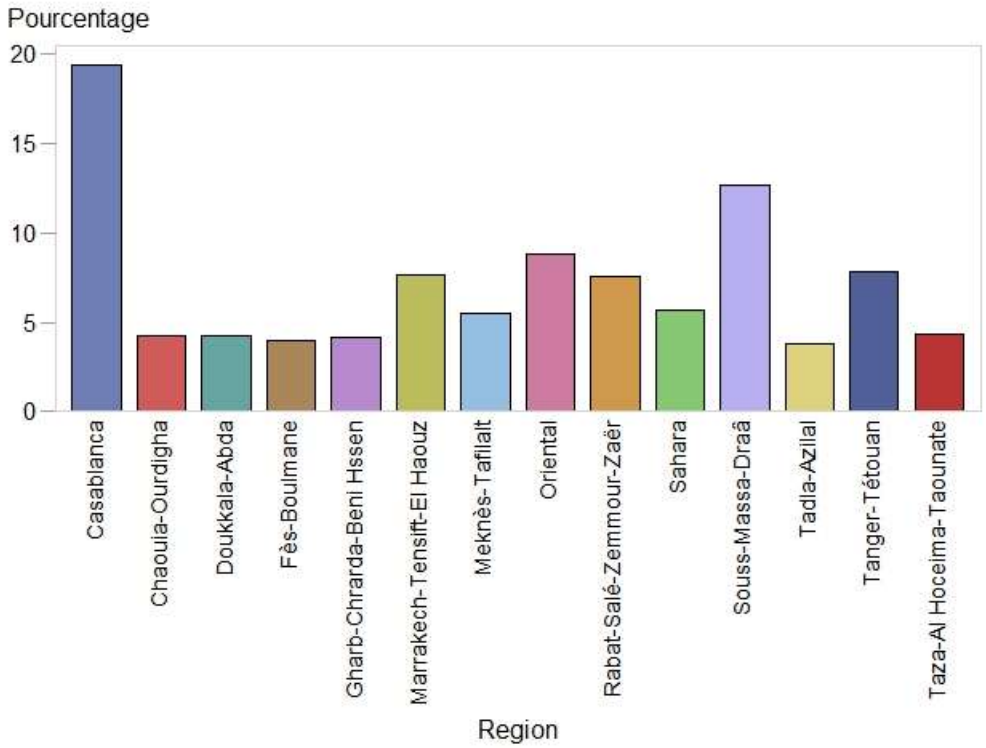


Figure 7: Les régions du portefeuille

Les figures ci-dessous représentent la région « SAHARA » ou les assurés renouvellent moins leurs contrats d'assurance automobile et la région « gharb chrarda beni hsen » qui est la région ou les assurés renouvellent plus leurs contrats.
Cette dernière est la région ou les assurés sont plus fidèles par rapport aux autres régions du royaume.

Procédure FREQ
Region=Sahara

renouvele	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
Non	33283	42.76	33283	42.76
Oui	44557	57.24	77840	100.00

Figure 8: Fréquence des assurés dans la région du « SAHARA »

Procédure FREQ

Region=Gharb-Chrarda-Beni Hssen

renouvele	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
Non	16193	28.48	16193	28.48
Oui	40663	71.52	56856	100.00

Figure 9: Fréquence des assurés dans la région « Gharb-Chrarda-Beni Hssen »

3.3 Coefficient de corrélation

Nous s'intéressons au test de corrélation de « Spearman » et de « Pearson »

Le coefficient de corrélation de rang (appelé coefficient de *Spearman*) examine s'il existe une relation entre le rang des observations pour deux caractères X et Y, ce qui permet de détecter l'existence de relations monotones (croissante ou décroissante), quelle que soit leur forme précise (linéaire, exponentiel, puissance, ...).

Ce coefficient est donc très utile lorsque l'analyse du nuage de point révèle une forme curviligne dans une relation qui semble mal s'ajuster à une droite.
Le calcul du coefficient de corrélation est une étape exploratoire qui doit être validé par un test de significativité qu'on détaillera dans la prochaine partie du mémoire.

Le coefficient de corrélation de « Pearson » examine s'il existe une relation linéaire.

La procédure « proc corr » est utilisée sous sas qui permet de calculer le coefficient de corrélation des variables quantitatives.
Le tableau ci-dessous est le résultat du calcul du coefficient de corrélation de *Spearman*. On remarque que la variable CRM est négativement corrélée à la variable cible du renouvellement.

Coefficients de corrélation de Spearman Proba > |r| sous H0: Rho=0 Nombre d'observations							
	age_permis	prime	CRM	COMMISSION	Profil_Risque	age_conducteur	age_vehicule
renouvele_b	0.06228	0.15604	-0.17971	0.12603	-0.09933	0.06316	-0.04321
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	1021107	1368133	1368133	1368133	1368133	1368133	1368073

Table 6: Coefficients de corrélation de Spearman

Le CRM, la prime et le profil du risque sont les variables les plus corrélées avec la variable « renouvele» selon le coefficient de corrélation de « *Spearman* ».

Par ailleurs, il est important de détecter la colinéarité entre variables explicatives. Nous s'intéressons aux liaisons qui peuvent exister entre les variables explicatives continues.

Le tableau ci-dessous fournit les corrélations entres les variables :

Coefficients de corrélation de Pearson Proba > |r| sous H0: Rho=0 Nombre d'observations									
	renouvele_b	AVENANT	COMMISSION	CRM	Profil_Risque	prime	age_conducteur	age_permis	age_vehicule
renouvele_b	1.00000	0.06410	0.11706	-0.15122	-0.09940	0.12689	0.05173	0.05411	-0.04241
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	1368133	1368133	1368133	1368133	1368133	1368133	1368133	1021107	1368073
AVENANT	0.06410	1.00000	0.00741	-0.08745	-0.05359	0.00478	0.05624	0.06019	0.02688
	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	1368133	1368133	1368133	1368133	1368133	1368133	1368133	1021107	1368073
COMMISSION	0.11706	0.00741	1.00000	-0.15023	-0.21227	0.83216	0.15589	0.19607	-0.31339
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	1368133	1368133	1368133	1368133	1368133	1368133	1368133	1021107	1368073
CRM	-0.15122	-0.08745	-0.15023	1.00000	0.47120	-0.14773	-0.16302	-0.15015	0.05497
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
	1368133	1368133	1368133	1368133	1368133	1368133	1368133	1021107	1368073
Profil_Risque	-0.09940	-0.05359	-0.21227	0.47120	1.00000	-0.21935	-0.31610	-0.34631	0.10050
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
	1368133	1368133	1368133	1368133	1368133	1368133	1368133	1021107	1368073
prime	0.12689	0.00478	0.83216	-0.14773	-0.21935	1.00000	0.16354	0.21503	-0.36103
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
	1368133	1368133	1368133	1368133	1368133	1368133	1368133	1021107	1368073
age_conducteur	0.05173	0.05624	0.15589	-0.16302	-0.31610	0.16354	1.00000	0.70897	-0.05018
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
	1368133	1368133	1368133	1368133	1368133	1368133	1368133	1021107	1368073
age_permis	0.05411	0.06019	0.19607	-0.15015	-0.34631	0.21503	0.70897	1.00000	-0.12349
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
	1021107	1021107	1021107	1021107	1021107	1021107	1021107	1021107	1021047
age_vehicule	-0.04241	0.02688	-0.31339	0.05497	0.10050	-0.36103	-0.05018	-0.12349	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
	1368073	1368073	1368073	1368073	1368073	1368073	1368073	1021047	1368073

Table 7: Corrélation croisées entre les variables

Cette étape est importante avant d'appliquer toute modélisation. Il faut s'assurer que les variables explicatives ne soient pas corrélées entre elle.

Cependant, on remarque que la variable « age_permis » et « age_conducteur » sont fortement corrélées de coefficient 0.709. Ainsi que la variable « CRM » et « Profil_Risque » de coefficient 0.471.

Etant donné qu'il est très difficile de travailler avec 165 variables et que l'efficacité d'un modèle se base sur la pertinence des variables et non leur quantité, nous essayerons d'en diminuer le nombre de façon à garder l'information transmise et donc assurer l'efficacité du modèle. Pour ne pas fausser le modèle, il ne faut pas introduire les variables qui sont corrélées entre elle.

L’analyse statistique a permis d’avoir une vision de la base de données. On a pu détecter les données manquantes et redondantes, valeurs aberrantes et d’analyser les variables qui influencent le renouvèlement du contrat d’assurance automobile

La prochaine partie du mémoire va traiter la modélisation de la variable cible « renouvele », en appliquant des modèles prédictifs qui sont l’arbre de décision, le modèle logistique et l’analyse discriminante.

III. Modèles prédictifs pour l'attrition

Dans cette section on va détailler le cadre théorique des modèles qu'on utilisera pour prédire le renouvellement du contrat d'assurance automobile. On s'intéresse à la génération du score d'attrition qui est la probabilité d'un client de quitter le portefeuille.

Les travaux menés dans ce sens sont plutôt développés en assurance vie « pour le rachat ».

1. Cadre théorique

1.1 Principe du Scoring

Le *scoring* se présente en effet comme un ensemble de méthodes conduisant à un classement d'individus au sein de groupes préalablement définis.

Formellement, étant donné un ensemble d'individus pouvant être décrits par un certain nombre de variables. Ces individus se répartissent entre quelques groupes définis à priori. Un individu se présente. On ne connait pas son groupe d'appartenance. Peut-on, sur la base des observations qu'il présente vis-à-vis des variables considérées, prévoir le groupe auquel il appartient ? C'est le problème auquel les méthodes de *scoring* cherchent à donner une solution.

Les principaux types de scores utilisés pour une compagnie d'assurance sont : le score d'appétence, le score de prime et le score d'attrition. Dans notre étude, on s'intéresse au *score d'attrition* qui est la probabilité qu'un client mette fin à son contrat d'assurance. Il est calculé pour un client de son contrat d'assurance depuis plusieurs mois sur la base de données constituée de ses garanties, du type de son contrat, de sa durée etc.

Le *scoring* est une méthode statistique permettant d'assigner une probabilité d'un évènement exacte à un assuré se présentant pour la première fois. Cette méthode permet donc d'estimer (du côté de l'assureur) les P_i du modèle précédent.

Sur la base de l'historique des contrats, la compagnie d'assurance évalue l'incidence de chacune des variables observables sur la probabilité de survenance de l'évènement.

Pour cela, on définit $Y_i$la variable indicatrice de renouvellement du contrat dans l'année pour l'individu i $Y_i = 1$ si l'assuré i a renouvelé son contrat, $Y_i = 0$ sinon et

X_i le vecteur de ses caractéristiques. On souhaite alors connaître l'effet de chacune des caractéristiques X_{it} sur Y_i.

Le caractère binaire de Y_i nous empêche cependant d'appliquer ici les méthodes de régressions linéaires classiques.

1.2 Limites du champ d'application du modèle classique

La famille des modèles linéaires classiques, bien que toujours largement employée aujourd'hui, présente cependant l'inconvénient d'avoir un champ d'application assez restreint.

En effet, certaines des hypothèses initiales ne peuvent être validées. Notre étude est un exemple, pour lesquels une telle modélisation n'est pas adaptée. On citera quelques hypothèses non vérifiées

- Hypothèse de normalité des résidus

La forme linéaire du modèle suppose la vérification, au moins a posteriori, de la distribution gaussienne de la variable cible, ce qui se valide par la normalité des résidus. Lorsque la variable à modéliser est une variable discrète ou se présente sous la forme de proportions ou de taux, cette hypothèse ne peut pas être validée.

- Hypothèse de variance constante

Il n'est pas rare de constater une augmentation de la variance à mesure que la moyenne augmente, ce qui ne peut donc pas être modélisé par un modèle linéaire classique.

1.3 Modèles prédictifs

Dans cette section, on va détailler le cadre théorique des modèles qu'on utilisera pour prédire le renouvellement du contrat d'assurance automobile.

Dans notre étude, on s'intéresse à la génération du *score d'attrition* qui est la probabilité qu'un client mette fin à son contrat d'assurance. Il est calculé pour un client de l'assurance depuis plusieurs mois sur la base de données constituée de ses garanties, du type de son contrat, de sa durée etc.

1.3.1 Analyse discriminante

Une technique très répondue pour décrire ainsi que pour prédire. L'analyse discriminante de Fisher fut parmi les grandes méthodes de classement.
Dans notre étude, on s'intéresse à l'analyse discriminante probabiliste. Elle se fonde sur l'approche dite bayésienne basant sur le théorème de Bayes

$$P(G_i/x) = \frac{P(G_i)P(x/G_i)}{\sum_j P(G_j)P(x/G_j)}$$

L'analyse discriminante est utilisée pour pouvoir construire une règle de décision.

Afin de mesurer le pouvoir discriminant de chaque variableX_j, il faut utiliser l'analyse de la variance à un facteur.
On considère 3 types de corrélation :

- La corrélation totale
- La corrélation intra classes
- La corrélation interclasse (les données sont résumées par les centres de gravité des classes pondérés par leurs fréquences)

Mesure de la qualité du modèle

Les mesures suivantes s'appliquent uniquement à l'analyse discriminante.

- Le lambda de *wilks*

Il s'agit du rapport $A = \frac{\det(W)}{\det(V)}$ du déterminant de la matrice des covariances intraclasse sur celui de la matrice des covariances totale

- Le coefficient de détermination R^2 (carré de la corrélation canonique)
- Le coefficient de détermination R^2 ajusté

$$R^2 ajusté = 1 - \frac{(1 - R^2) \times (n - 1)}{n - p - 1}$$

Le R^2ajusté peut être négatif et est toujours $< R^2$

Mesures d'efficacité :

Les trois statistiques suivantes permettent de tester globalement la qualité d'ajustement du modèle aux données. Plus leur valeur est petite, meilleur est le modèle.

Soient m le nombre de paramètres, L la vraisemblance et n le nombre d'observations

- -2Log(L)

Cette statistique est importante quand la vraisemblance est faible

- Critère d'information d'Akaïke AIC=-2 Log L+2m

Ce critère prend en compte le nombre de variables explicatives. Deux modèles ayant la même valeur pour le critère précédent, ne seront pas jugés équivalents pour ce critère si leurs nombres de paramètres soient différents. Celui qui a le moins de variables explicatives sera préféré.

- Critère de Schwartz SC= -2LogL+m Log n

Ce critère prend donc en compte le nombre d'observations.

La partie théorique sur la régression logistique et l'arbre de décision a été rédigée à partir du livre de Tufféry Stéphane « Data Mining et statistique décisionnelle ».

1.3.2 *Régression logistique*

La régression logistique est une technique de modélisation qui est très répondue, elle est un cas particulier du modèle linéaire généralisé.
Elle vise à prédire et expliquer les valeurs d'une variable à partir de variables explicatives.

On s'intéresse ici à la régression logistique binaire. Elle consiste à considérer une variable cible binaire Y=0 ou Y= 1, et p variables explicatives.

L'objectif de cette méthode est celle de toute régression, modéliser l'espérance conditionnelle

$$E(Y/(X = x))$$

On cherche la valeur moyenne de Y pour toute valeur de X. Pour une valeur Y valant 0 ou 1 (loi de Bernoulli), cette valeur moyenne est la probabilité que Y = 1

$$E(Y/(X = x)) = P(Y = 1/(X = x))$$

Généralement la fonction utilisée est la fonction *logit* qui est une bijection de $]0,1[$ à R définie ci-dessous :

$$logit(s) = \ln\left(\frac{s}{1-s}\right)$$

La formule du modèle logistique est

$$\ln\left(\frac{P\big((Y_i = 1)\big)}{P(Y_i = 0)}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

On a $P(Y_i = 1) = 1 - P(Y_i = 0)$

Cependant

$$P(Yi = 0) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}$$

D'où la formule obtenue

$$P(Y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}$$

Le rapport $\frac{P(Y_i=1)}{P(Y_i=0)}$ exprime un odds (côtes lors d'un pari). Par exemple si la probabilité qu'un assuré renouvelle son contrat est de 0.6 et la probabilité qu'il ne

renouvelle pas son contrat est de 0.2, le odds-ratio vaut 0.6/0.2=3. Il y a 3 fois plus de chance qu'un assuré renouvelle son contrat que de ne pas le faire.

Les observations des m variables explicatives sont organisées dans la matrice *X*, de dimension$(\mathrm{n} \times \mathrm{m})$. Les p coefficients de régression s'inscrivant dans un vecteur βde dimension$(\mathrm{m} \times 1)$, nous pouvons définir le prédicteur linéaire η , composante déterministe du modèle $\eta = \mathrm{X} \times \beta$

La fonction de lien exprime le lien existant entre η et l'espérance de Y.
Rappelons que celle-ci doit, entre autres, être une fonction monotone et dérivable.

1.3.2.1 Estimation des coefficients

L'estimation des paramètres β est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé.

- Equations de vraisemblance :

La vraisemblance s'écrit :

$$L(\beta_0 + \cdots + \beta_m) = \prod_{i=1}^{n} \frac{exp(y_i \sum_{k=0}^{m} \beta_k x_{ik})}{1 + \exp(y_i \sum_{k=0}^{m} \beta_k x_{ik})}$$

Les paramètres $\widehat{\beta_k}$sont alors des solutions des $m + 1$ équations suivantes :

$$\sum_{i=1}^{n} \left(y_i x_{ik} - \frac{x_{ik} exp(y_i \sum_{k=0}^{m} \beta_k x_{ik})}{1 + exp(\sum_{k=0}^{m} \beta_k x_{ik})} \right) = 0 \qquad \forall\, k = 0, \dots, m$$

- Résolution du système

Ces équations, non linéaires, n'ont pas de solution analytique : leur résolution requiert l'utilisation de méthodes itératives, dans lesquelles interviennent le Hessien (algorithme de Newton-Raphson), ou la matrice d'information de Fisher (méthodes des scores de Fisher).

1.3.3 Arbre de décision

Les arbres de décision sont utilisés pour prévoir l'affectation d'observations ou d'objets à des classes de variables dépendantes catégorielles à partir de leurs mesures sur une ou plusieurs variables prédictives. La flexibilité des arbres de décision est en fait une analyse très attrayante car elle ne dépend pas de la distribution des données. Elle peut donc être utilisée comme une technique exploratoire.

Dans notre étude, on l'utilise pour classifier les variables explicatives ainsi que pour prédire le comportement de renouvellement du contrat d'assurance automobile.

L'objectif est de construire des sous-groupes les plus homogènes du point de vue de la variable à prédire « renouvele »

Les arbres de décision sont très efficaces pour des tailles d'échantillon importantes et facile à implémenter comme elles ne requièrent pas d'hypothèses sur la distribution des variables.

Les importants algorithmes d'apprentissage par arbres de décision sont ceux développés par Leo Breiman[2], sa proposition vise à corriger plusieurs inconvénients connus de la méthode initiale.

Les arbres de décision sont utilisés dans le cadre de la découverte de connaissances dirigée. Ce sont des outils très puissants principalement utilisés pour la classification, la description ou l'estimation.

Le principe de fonctionnement est le suivant : pour expliquer une variable, le système recherche le critère le plus déterminant et découpe la population en sous populations possédant la même entité de ce critère. Chaque sous population est ensuite analysée comme la population initiale.

Le modèle rendu est facile à comprendre et les règles trouvées sont très explicites. Le but de cette technique est de créer un arbre de décision procédant à une analyse critère par critère. La détermination de ces critères significatifs est faite selon les poids statistiques des valeurs.

L'outil de *data Mining* va parcourir les différents critères possibles, dont la finalité sera de trouver des liens entre les chemins qui ont une signification par rapport à la problématique donnée.

On donne un ensemble X de Ndont les éléments sont notés $x_i$et dont les P attributs sont quantitatifs. Chaque élément de X est étiqueté, c'est-à-dire qu'il lui est associé une classe ou un attribut cible que l'on note y appartenant à Y.
A partir de ce qui précède, on construit un arbre dit « de décision » tel que :

- Chaque nœud correspond à un test sur la valeur d'un ou plusieurs attributs.
- Chaque branche partant d'un nœud correspond à une ou plusieurs valeurs de ce test.
- Nœuds feuilles correspondent à un classement.

[2] Leo Breiman , sa contribution la plus importante concerne son travail sur les arbres de régression et de classification et les ensembles d'arbres taillés pour les échantillons traités par les techniques de bootstrap

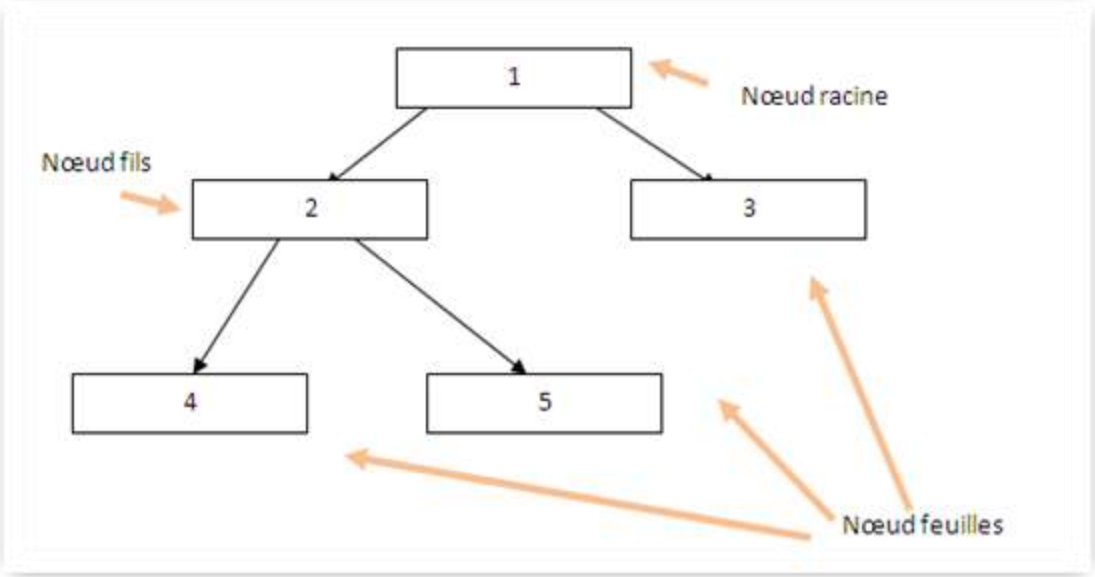


Figure 10: Arbre de décision

Les principaux algorithmes d'arbres de décision sont :

- CART (*Classification And Regression Tree*), qui est adapté avec tout type de variables.
- CHAID (*Chi-Square Automation Interaction Detection*), initialement réservé à l'étude des variables discrètes et qualitatives.

Algorithme de la méthode CHAID :

1. On construit pour chaque prédicteur X_i, le tableau de contingence $X_i \times Y$ et on effectue les étapes 2 et 3.

2. On sélectionne la paire de modalités de X_i dont le sous-tableau $(2 \times k)$ a le plus petit χ^2. Si ce χ^2 n'est pas significatif, on fusionne les 2 modalités et on répète cette étape.

3. Eventuellement, pour chaque modalité composée de plus de 3 modalités originales, on détermine la division binaire au χ^2 le plus grand. S'il est significatif, on effectue cette division.

4. On calcule la significativité (probabilité associée au χ^2) de chaque prédicteur X_i dont les modalités ont été précédemment regroupées et on retient le plus significatif. Si ce χ^2 est plus significatif que le seuil choisi, on peut diviser le nœud

en autant de nœuds-fils qu'il y a de modalités après regroupement. Si ce χ^2n'atteint pas le seuil spécifié, le nœud n'est pas divisé.

5. Le critère d'arrêt de l'arbre de décision dépend du type et du paramétrage de l'arbre.

Souvent combine plusieurs règles :

- La profondeur de l'arbre a atteint une limite fixée
- Ou le nombre de feuilles (c'est-à-dire de règles) a atteint un maximum fixé
- Ou l'effectif de chaque nœud est inférieur à une valeur fixée en deçà de laquelle on estime qu'il ne faut plus diviser un nœud (au moins 75 à 100 pour de bons résultats)
- Ou la division ultérieure de tout nœud provoquerait la naissance d'un fils d'effectif inférieur à une valeur fixée
- Ou la qualité de l'arbre est suffisante
- Ou la qualité de l'arbre n'augmente plus de façon sensible.

Avantages de la méthode Chaid :

L'un des avantages de cette méthode est la discrétisation automatique des variables continues.

Dans notre cas, on cherche à prédire une variable cible « renouvele » à l'aide de plusieurs variables explicatives que nous voulions découper en classes pour les raisons déjà indiquées alors cette méthode semble la plus adéquate et on le fera en utilisant le logiciel IBM SPSS.

Algorithme de la méthode CRT:

L'arbre CART ou CRT inventé en 1984 par les statisticiens L.breiman, J.H Frriedman, R.A. Olshen et C.J. Stone est l'un des plus performants et des plus répandus. Cet algorithme peut servir au classement comme à la prédiction. CART court à l'indice de Gini pour trouver la meilleure séparation de chaque nœud.

Un arbre est construit, puis l'algorithme en déduit plusieurs sous arbres par élagages successifs, qu'il compare, avant de retenir celui pour lequel le taux d'erreur mesuré est le plus bas possible. Il traite aussi les valeurs manquantes.

Critère de séparation :

Le choix de la meilleure séparation d'un nœud s'appuie sur des critères, les plus répondus sont les suivants :

- χ^2 : quand les variables explicatives sont qualitatives ou discrètes.

Gini : pour tout type de variables explicatives (utilisé dans l'arbre CART).Le premier critère est bien connu. Le choix de la variable explicative et les modalités de cette variable devant séparer un nœud en plusieurs nœuds fils est fait en sorte de maximiser la valeur de χ^2de cette variable croisée avec la variable cible.

L'indice de Gini d'un nœud est calculé ainsi :

$$Gini(noeud) = 1 - \sum_i {f_i}^2 = \sum_{i \neq j} f_i f_j$$

où les f_i , $i = 1 \text{ à } p$, sont les fréquences relatives dans le nœud des p classes à prédire(variable cible).
Plus les classes sont uniformément distribuées dans un nœud, plus l'indice de Gini est élevé.

L'indice de Gini mesure la probabilité que deux individus, choisis aléatoirement (avec remise) dans un nœud, appartiennent à deux classes différentes. Il permet de prendre en compte les coûts de mauvaise affectation C_{ij} d'un individu de la classe j dans la classe i. il est défini comme :

$$Gini(noeud) = \sum_{i \neq j} C_{ij} f_i f_j$$

Avantage des arbres de décision

- Aucune condition sur les variables explicatives (l'arbre ne suppose pas des lois probabilistes particulières sur les variables explicatives)
- Certains arbres comme CART permettent de traiter tous types de variables : discrètes, continues, catégorielles et de gérer judicieusement les données manquantes.
- Simple à utiliser et une grande rapidité d'exécution.

1.4 Comparaison des modèles

1.4.1 Matrice de confusion

La régression logistique peut se révéler très utile lorsque nous souhaitons classer les clients selon leurs degrés de fidélité ou introduire d'autres calculs ultérieurement.

Les méthodes d'évaluation sont basées sur les probabilités à postériori trouvées par le modèle tout en s'appuyant sur le calcul de la matrice de confusion.

Cette matrice confronte toujours les valeurs observées de la variable cible « renouvele » dans notre étude, avec celles qui sont prédites, puis comptabilise les bonnes et les mauvaises prédictions. Son intérêt est qu'elle permet à la fois d'appréhender le taux d'erreur et de rendre compte de la structure de l'erreur.

A partir de la forme générique de la matrice de confusion, plusieurs indicateurs peuvent être déduits pour rendre compte de la concordance entre les valeurs observées et les valeurs prédites de la variable cible. Ces indices fournissent des informations très intéressantes concernant notre variable :

- VP sont les vrais positifs, c'est-à-dire les observations qui ont été classées positives et qui le sont réellement.
- FP sont les faux positifs, c'est-à-dire les individus classés positifs et qui sont en réalité des négatifs.
- De la même manière, les FN sont les faux négatifs et VN sont les vrais négatifs.

Sauf que ces termes sont peu utilisés en pratique car les positifs et les négatifs n'ont pas le même statut dans la majorité des études.

La sensibilité ou le taux de vrais positifs TVP indique la capacité du modèle à retrouver les positifs. Elle se calcule comme suit :

$$Sensibilité = \frac{VP}{VP + FP}$$

La spécificité représente le taux de faux contrats renouvelés

$$\begin{aligned} Spécificité &= 1 - TFP \\ &= 1 - \frac{FP}{VP+FP} \end{aligned}$$

Le taux de bonne détection global correspond à la probabilité de bon classement du modèle, il se calcule comme suit :

$$Taux\ de\ bonne\ détection = \frac{VP + VN}{Total}$$

La sensibilité et la spécificité jouent un rôle particulier dans l'évaluation du modèle. Un « bon » modèle doit présenter des valeurs assez fortes de taux de bonne détection, des valeurs élevées de sensibilité, précision et spécificité.

1.4.2 Courbe ROC

Les notions expliquées ci-dessus sont employées pour caractériser le modèle et tracer la courbe de ROC. Elle se construit en représentant sur l'axe des abscisses le taux des faux contrats renouvelés, et en ordonnées le taux des vrais contrats renouvelés, avec un seuil de sensibilité et de spécificité variable et qui varie entre 0 et 1.

On peut ainsi visualiser le pouvoir discriminant d'un modèle de score à l'aide de la courbe de ROC. En effet, si cette courbe coïncide avec la diagonale, c'est que le modèle n'est pas plus performant qu'un modèle aléatoire, par contre, plus cette courbe est proche du coin du carré, plus le modèle sera meilleur, du fait qu'il permet de capturer davantage de vrais événements avec le moins possible de faux événements. Ci-dessous une représentation de la courbe ROC.

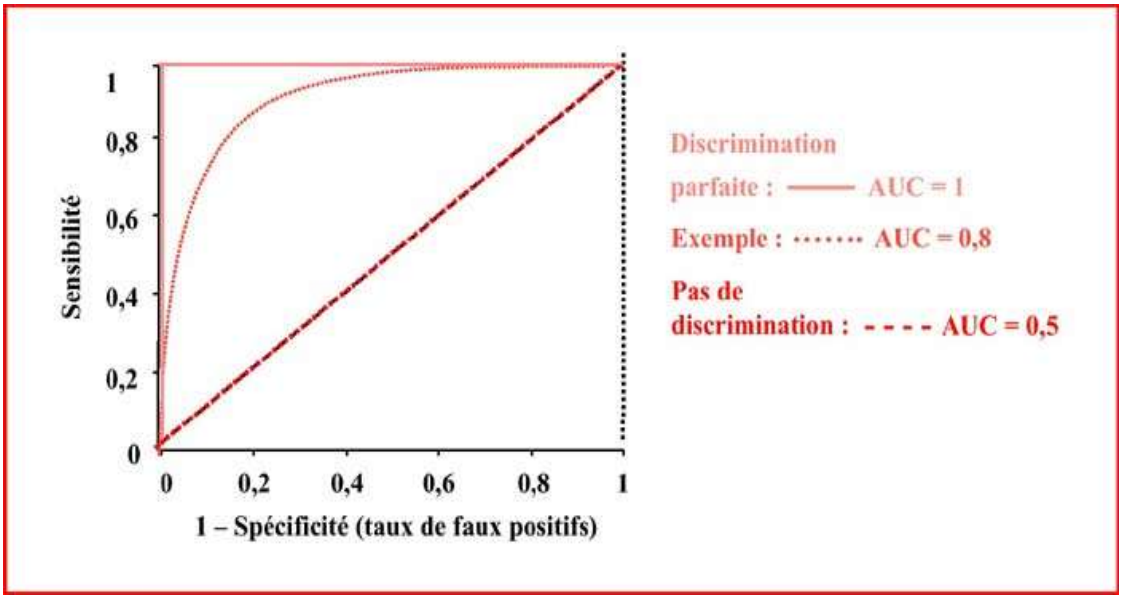


Figure 11: Représentation de la courbe ROC

La surface située sous la courbe est le critère AUC. Elle exprime la probabilité de placer un individu positif devant un négatif. Ainsi, dans le cas d'une discrimination parfaite, les positifs sont sûrs d'être placés devant les négatifs, nous avons AUC = 1. Quand la courbe ROC se confond avec la première bissectrice, nous avons AUC = 0,5.

2. Sélection des variables

2.1 Différentes méthodes

Lors de l'étude exploratoire, nous avons eu une idée sur les variables susceptibles d'expliquer l'acte de renouvellement du contrat de la part de l'assuré.

Il n'est cependant pas très pertinent de conserver autant de variables dans la modélisation que nous utiliserons pour prédire la probabilité de renouvèlement. Il nous faut donc sélectionner certaines variables que nous jugeons importantes pour l'étude. Pour cela, nous disposons avec le logiciel SAS de 3 méthodes de sélection :

- Méthode ascendante « *Forward* »

A la première étape, seule la constante est introduite dans le modèle. Puis, à chaque étape suivante, la variable la plus significative selon le critère du chi deux résiduel, c'est-à-dire qui permet d'expliquer le plus la variable cible, parmi celles qui restent, est introduite dans le modèle. Cette procédure se termine lorsqu'il n'y a plus de variable suffisamment significative suivant le seuil que l'on s'est fixé.

- Méthode descendante « *Backward* »

A la première étape, toutes les variables sont introduites dans le modèle. Puis, à chaque étape suivante, la variable la moins significative selon le test du chi deux de Wald est retirée de la modélisation. A nouveau, cette procédure se termine lorsque toutes les variables restant dans le modèle sont significatives selon un seuil fixé d'avance.

- Méthode pas à pas « *Stepwise* »

Cette méthode est une combinaison des deux précédentes. La première étape consiste à modéliser la variable cible en introduisant uniquement la constante. A

chaque étape suivante, comme pour la méthode ascendante, la variable la plus significative est introduite dans le modèle. Dans le même temps, les variables qui constituent le modèle sont testées et, si l'une d'elles n'est plus significative à la suite de l'introduction d'une autre variable, elle est retirée du modèle.

2.2 La méthode utilisée

Nous avons testé les trois méthodes de sélection. La représente le choix des variables pour chaque méthode.

	Méthode Forward	**Méthode Backward**	**Méthode Stepwise**
1	CRM	CRM	CRM
2	Prime	Début	Prime
3	Puissance_fiscale	Ind_incendie	Puissance_fiscale
4	Ind_incendie	Puissance_fiscale	Age_vehicule
5	Ind_vol	Age_vehicule	Ind_incendie
6	Ind_dommage	Age_conducteur	Ind_vol

Figure 12: Récapitulatif des choix des variables selon chaque méthode de sélection

Nous constatons que presque les mêmes variables sont retenues pour la méthode ascendante et la méthode pas à pas.
La méthodologie utilisée est de tester les 3 méthodes.

La méthodologie appliquée est de faire une sélection des variables les plus discriminantes par rapport à la variable cible « renouvele ».

La méthode de sélection ascendante « *forward* » repose sur les étapes suivantes :

- Introduite au début la constante seule dans le modèle.
- A chaque étape suivante, la variable la plus significative selon le critère du chi deux résiduel, c'est-à-dire qui permet d'expliquer le plus la variable cible, parmi celles qui restent, est introduite dans le modèle.

- Cette procédure se termine lorsqu'il n'y a plus de variable suffisamment significative suivant le seuil que l'on s'est fixé.

L'ordre d'entrée des critères explicatifs dans le modèle est le suivant :
Les résultats sont présentés dans le tableau ci-dessous :

Forward Selection Summary								
Etape	Saisi	R carré partiel	Valeur F	Pr > F	Lambda de Wilk	Pr < Lambda	Corrélation canonique moyenne au carré	Pr > ASCC
1	CRM	0.0246	15876.0	<.0001	0.97543480	<.0001	0.02456520	<.0001
2	prime	0.0114	7240.10	<.0001	0.96435925	<.0001	0.03564075	<.0001
3	PUISSANCE_FISCALE	0.0035	2199.66	<.0001	0.96100602	<.0001	0.03899398	<.0001
4	Ind_Incendie	0.0003	217.94	<.0001	0.95986100	<.0001	0.04013900	<.0001
5	Ind_Vol	0.0006	372.08	<.0001	0.95929479	<.0001	0.04070521	<.0001
6	Indic_Dommage	0.0004	235.48	<.0001	0.95893659	<.0001	0.04106341	<.0001
7	Ind_Pro_Cond_Pass	0.0003	208.95	<.0001	0.95861885	<.0001	0.04138115	<.0001
8	age_conducteur	0.0002	115.50	<.0001	0.95844324	<.0001	0.04155676	<.0001
9	age_permis	0.0003	216.10	<.0001	0.95811481	<.0001	0.04188519	<.0001
10	ancienneté_permis	0.0001	55.85	<.0001	0.95802994	<.0001	0.04197006	<.0001
11	age_vehicule	0.0002	120.61	<.0001	0.95784668	<.0001	0.04215332	<.0001
12	Ind_Pro_Juri	0.0001	46.54	<.0001	0.95777597	<.0001	0.04222403	<.0001
13	Ind_Collision	0.0001	37.15	<.0001	0.95771953	<.0001	0.04228047	<.0001
14	Ind_DTR	0.0001	37.01	<.0001	0.95766330	<.0001	0.04233670	<.0001
15	Code_PDV	0.0000	29.41	<.0001	0.95761863	<.0001	0.04238137	<.0001
16	VALEUR_VENALE	0.0000	25.62	<.0001	0.95757971	<.0001	0.04242029	<.0001
17	VALEUR_DES_GLACES	0.0000	18.52	<.0001	0.95755158	<.0001	0.04244842	<.0001
18	Ind_Inondation	0.0000	17.51	<.0001	0.95752498	<.0001	0.04247502	<.0001
19	Profil_Risque	0.0000	16.98	<.0001	0.95749919	<.0001	0.04250081	<.0001
20	Ind_Perte_Fin	0.0000	11.61	0.0007	0.95748156	<.0001	0.04251844	<.0001

Figure 13: Résultats de la sélection "Forward"

La Figure 13 représente l'ordre de sélection des variables, le système s'est arrêté à l'étape 21 lorsqu'il n'y a plus de variable suffisamment significative puisqu'il estime que plus aucune variable explicative ne doit entrer.

Nous constatons que les variables "Ind_Perte_Fin" et "Profil_Risque" sont peu discriminantes (19 et 20 ème positions) pour cette méthode de sélection.

Comme nous l'avons constaté au chapitre précèdent lors du l'analyse uni-variée, les variables "CRM" et "prime" sont les plus corrélées à la variable «renouvele».

3. Réalisation d'un échantillonnage

Avant de passer à la modélisation, nous procédons à l'échantillonnage.
Cependant, l'échantillonnage n'est réalisé que si les données sont suffisamment importantes qui est le cas dans notre étude. Ils existent plusieurs techniques d'échantillonnage.

Les principaux techniques sont l'échantillonnage simple, l'échantillonnage stratifié et l'échantillonnage systématique.

- L'échantillonnage simple consiste à tirer au hasard des individus dans une population selon une loi statistique que l'on peut choisir (normale, uniforme, exponentielle…). Chaque individu ayant la même probabilité d'être tiré.

- L'échantillonnage stratifié consiste à réaliser une partition de la population par exemple une répartition des clients en tranches d'âges. A tirer les clients au sort dans chaque strate de la partition de façon à obtenir un sous échantillon par strate, puis à réunir tous ces échantillons

3.1 Pourquoi avoir recours à un échantillonnage ?

L'échantillonnage constitue une étape fondamentale pour élaborer un modèle prédictif. Il est le cœur de la problématique en assurance automobile.

Les techniques de modélisation travaillent sur deux échantillons de la population :

• L'échantillon d'apprentissage.
• L'échantillon de test.

L'échantillon d'apprentissage est celui avec lequel le modèle est construit.

L'échantillon de test est celui avec lequel le modèle est testé. Ces échantillons doivent être représentatifs pour garantir la qualité du modèle.

L'échantillon d'apprentissage est en général très grand et représente plus de 70% de la base de données afin d'avoir les meilleurs résultats possibles et de donner une meilleure représentation de la base de données, tandis que l'échantillon test est plus petit et sert à valider le modèle retenu et d'affirmer sa justesse.

L'échantillonnage est une étape importante de toute technique de data mining. La majorité des algorithmes mettent en œuvre un échantillon d'apprentissage pour appliquer le modèle et un échantillon test pour valider le modèle.

3.2 L'échantillonnage adopté pour l'étude

Dans notre étude, nous optons pour la technique classique d'un échantillonnage simple. Ce choix de cette méthode nous permet d'avoir la même proportion des contrats d'assurance non renouvelés dans la base d'apprentissage et la base test.

La part de ceux non renouvelés (ceux qui ont résilié leurs contrats à l'échéance) dans la base initiale est relativement faible (autour de 32,45% de la base étudiée).

Sur la base initiale, nous réalisons un tirage aléatoire sans remise, basé sur une loi uniforme (en gardant l'homogénéité de la base initiale) afin d'obtenir une part de 80% « échantillon d'apprentissage » et le second échantillon test permet quant à lui d'analyser la solidité du modèle réalisé.
La variable « renouvele_b » est une variable binaire, elle prend 1 si l'assuré a renouvelé son contrat ou 0 sinon.

Les images ci-dessous représentent :

- La base totale est la base de départ. Elle est constituée de 1368133 expositions contrats d'assurance automobile.
- La base d'apprentissage : elle est construite par tirage de 1094496 individus en s'appuyant sur la loi uniforme.
- La base test ce qui reste après le tirage. Elle représente 20% de la base initiale. Nous utilisons une macro sur SAS qui permet de tirer aléatoirement 80% de la base initiale, soit 1094496 sans remise avec une loi uniforme.

  Nous optons pour la technique de validation croisée « *cross-validation* » appelée «*testset validation*».

```
%macro tasr(libref=,entree=,sortie=,nb=);
data &libref..&sortie (drop=i j count);
  count=0;
  array obsnum(&nb) _temporary_;
  do i=1 to &nb;
      redo:
      select=ceil(ranuni(100)*n);
      set &libref..&entree point=select nobs=n;
         do j=1 to count;
           if obsnum(j)=select then goto redo;
         end;
      position=select;
      count=count+1;
      obsnum(count)=select;
      output;
   end;
   stop;
   set &libref..&entree;
run;

                                %mend;
```

Figure 14: Macro utilisée sur SAS pour la méthode d'échantillonnage "cross-validation"

L'avantage de cette méthode est qu'elle est simple et facile à utiliser. Le seul inconvénient est qu'elle prend du temps quand le nombre à générer est grand.

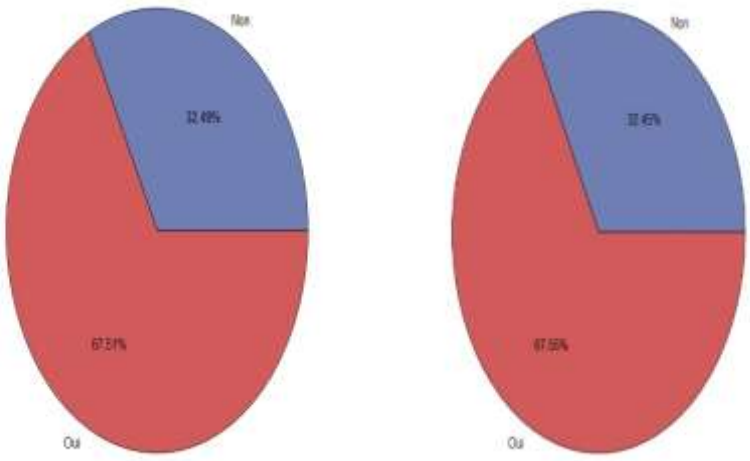


Figure 15: Représentation du pourcentage de la variable « renouvele » sur la base d'apprentissage et base test

Ces échantillons doivent être représentatifs pour garantir la qualité du modèle. Nous avons choisis cette méthode d'échantillonnage car elle permet de garder la même proportion dans chaque base.

4. Application : Modélisation

Dans cette partie, les résultats de l'application des modèles sont présentés.
Les logiciels utilisés sont SAS et IBM SPSS.

L'élaboration des modèles prédictifs constitue le cœur de l'activité de datamining. Vis-à-vis de notre objectif cette étape consiste en un calcul d'un score et la construction d'un classifieur. Les principales méthodes de *scoring* sont l'analyse discriminante, la régression logistique et les arbres de décision.

Ce chapitre propose, dans un premier temps, de classifier les variables continues les plus discriminantes, par la suite réaliser une analyse discriminante binaire afin de définir une fonction de résiliation permettant d'appréhender, de manière mathématique, le comportement des clients vis à vis de la conservation de leur contrat automobile.
Après avoir réalisé une analyse discriminante et interprété les résultats qui en découlent, nous appliquons la régression logistique et l'arbre de décision pour prédire le comportement du renouvellement d'un contrat d'assurance automobile.

4.1 Classification de variables continues

Dans l'approche par segmentation, la population de départ est découpée de manière successive selon les modalités des variables déterminées comme les plus discriminantes à chaque itération. Différents algorithmes de segmentation existent, tel que l'algorithme CHAID d'un arbre de décision.

Les modèles arbres de décision sont appliqués dans deux approches différentes :

- La première approche est de classifier quelques variables quantitatives, cela nous a permis d'améliorer le modèle de régression logistique et de mieux interpréter les résultats.

- La deuxième approche est de prédire le comportement du renouvellement de l'assuré de son contrat d'assurance automobile « partie 4.4 »

Mise en pratique des arbres de décision sur IBM SPSS Statistics en s'appuyant sur l'algorithme CHAID.

Nous avons vu dans la partie théorique de la construction d'un arbre que l'algorithme CHAID permet de découper en k classes.

Les figures ci-dessous montrent les règles de décisions qui consistent à la construction de la variable « Tranche_crm » et la variable « Tranche_profil_risque ».

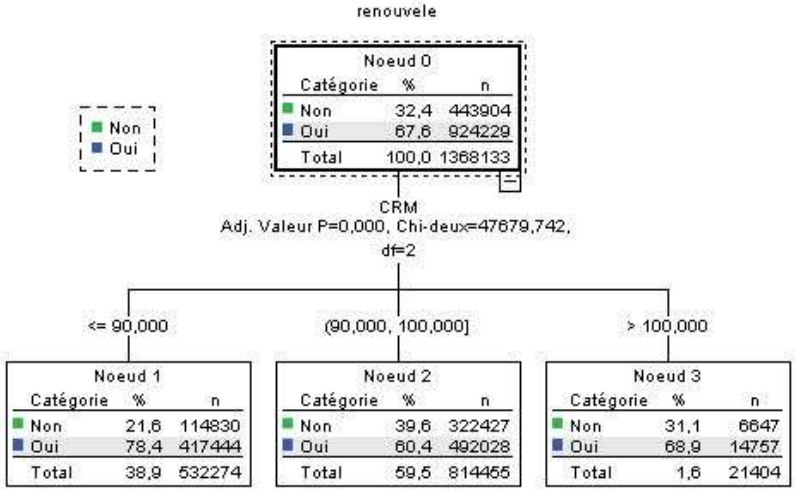


Figure 16: Discrétisation de la variable "CRM" par CHAID

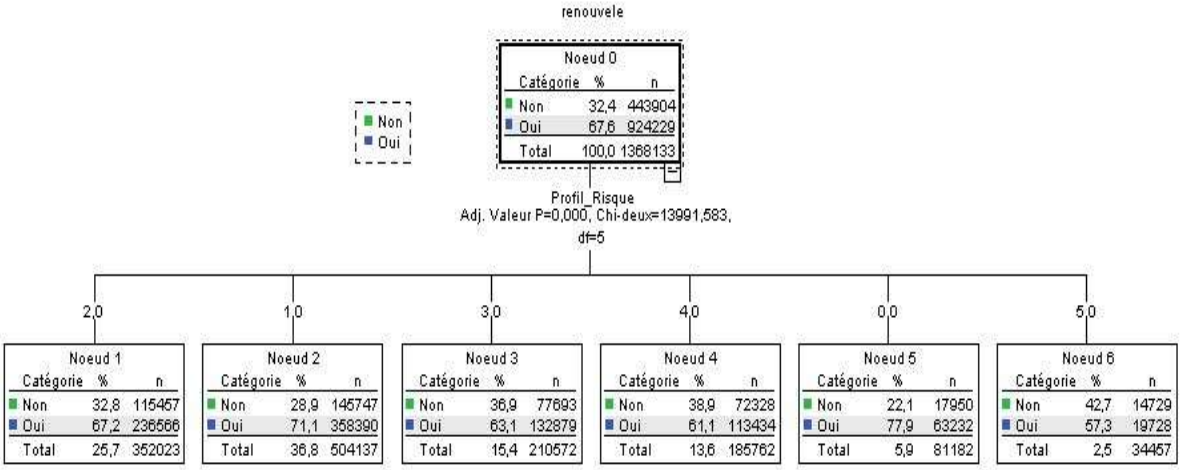


Figure 17: Discrétisation de la variable "Profil_Risque" par CHAID

La variable « Profil_Risque » est découpée en 6 nœuds. Nous remarquons que les assurés les plus risqué du profil de risque égal à 5 sont les plus à ne pas renouveler leurs contrats.

Le même raisonnement est adopté pour les autres variables de classification voir Annexes.

4.2. Analyse discriminante

L'analyse discriminante est une technique descriptive et aussi prédictive. Selon les objectifs on distingue ces deux approches.
Elle est utilisée pour déterminer les variables qui permettent de discriminer deux ou plusieurs groupes se produisant naturellement.

La procédure utilisée pour appliquer l'analyse discriminante est « proc discrim » de SAS qui permet de prédire. Elle s'applique uniquement sur les variables quantitatives.

Elle s'appuie sur la méthode bayésienne qui consiste à calculer, à posteriori, les probabilités d'appartenance à un groupe. Ci-dessous les résultats obtenus :

Le Système
SAS

The DISCRIM Procedure

Number of Observations and Percent Classified into renouvele_2

De renouvele_2	0	1	Total
0	424817	6198	431015
	98.56	1.44	100.00
1	195674	3717	199391
	98.14	1.86	100.00
Total	620491	9915	630406
	98.43	1.57	100.00
Priors	0.68371	0.31629	

Error Count Estimates for renouvele_2

	0	1	Total
Rate	0.0144	0.9814	0.3202
Priors	0.6837	0.3163	

Figure 18: Matrice de confusion de l'analyse discriminante

L'erreur du modèle analyse discriminante sur les variables quantitatives est de 31.63%.

Multivariate Statistics

Statistique	Valeur	Valeur F	DDL Num.	DDL Res.	Pr > F
Wilks' Lambda	0.957482	1333.01	21	630384	<.0001
Pillai's Trace	0.042518	1333.01	21	630384	<.0001
Average Squared Canonical Correlation	0.042518				

Figure 19: Statistique F du modèle analyse discriminante

Les variables non retenues de l'analyse discriminante sont présentées dans le tableau ci-dessous :

Statistics for Entry, DF = 1, 630383				
Variable	R carré partiel	Valeur F	Pr > F	Tolérance
MONTANT_ACCESSOIRES	0.0000	4.33	0.0374	0.0012
POIDS_EN_CHARGE	0.0000	6.31	0.0120	0.0012
Valeur_Conventionnelle	0.0000	4.51	0.0336	0.0012
VALEUR_NEUVE	0.0000	4.46	0.0348	0.0012
Ind_BDG	0.0000	0.99	0.3200	0.0012
Ind_Tierce	0.0000	1.09	0.2962	0.0006
Ind_Coll_Etendue	0.0000	0.00	0.9552	0.0012
Ind_Tierce_Limite	0.0000	0.13	0.7189	0.0012

Figure 20: Variables non retenues de l'analyse discriminante

Voir annexe l'ensemble des résultats de l'analyse discriminante.

4.3. Modèle 1 : Régression logistique, analyse discriminante et arbre de décision

La régression logistique est la méthode la plus fiable des méthodes de classification binaire.

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapport de vrais	60973.0513	37	<.0001
Score	59200.9091	37	<.0001
Wald	56083.6693	37	<.0001

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > Khi-2
Tranche_prime	9	16598.8753	<.0001
puiss_fisc	4	3545.5653	<.0001
age_cond	7	1367.5535	<.0001
tranche_crm	2	17362.7220	<.0001
COMBUSTION	1	800.5574	<.0001
Contentieux	1	525.7653	<.0001
age_veh	8	739.4309	<.0001
Difficulte	1	348.9171	<.0001
Profil_Risque	5	973.7444	<.0001

Figure 21: Impact global des variables explicatives du modèle « régression logistique »

Les variables retenues sont toutes significatives de P-value< 0.0001.
On peut noter l'importance de quelques variables telles que : « tranche_crm », « Tranche_prime » ou encore « puiss_fisc ».

Un extrait du tableau « Estimation par l'analyse du maximum de vraisemblance » est le plus utile pour le scoring, puisqu'il contient les coefficients de chaque variable dans la régression logistique (ou plus précisément les estimateurs de ces coefficients). Il s'agit de la colonne « Estimation ». Chaque estimateur est soumis à une certaine incertitude mesurée par son écart-type (« Erreur std ») et le ratio $\left(\frac{Estimation}{Erreur}\right)^2$ est d'autant plus élevé que l'estimateur est significativement différent de 0.

Estimations par l'analyse du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		1	0.4143	0.0277	223.9483	<.0001
Tranche_prime	11761.164 <prime <=13067.966	1	0.8912	0.0112	6297.9776	<.0001
Tranche_prime	13067.966 <prime <=16898.580	1	1.0515	0.0103	10404.9184	<.0001
Tranche_prime	16898.580 <prime	1	1.1697	0.0116	10200.4238	<.0001
Tranche_prime	2672.784 <prime <=3175.545	1	0.3915	0.00994	1551.1955	<.0001
Tranche_prime	3175.545 <prime <=3528.312	1	0.5244	0.0114	2109.0370	<.0001
Tranche_prime	3528.312 <prime <=3593.732	1	0.7470	0.00962	6027.8509	<.0001
Tranche_prime	3593.732 <prime <=5069.520	1	0.7560	0.0114	4383.8779	<.0001
Tranche_prime	5069.520 <prime <=6759.432	1	0.8116	0.00996	6643.8527	<.0001
Tranche_prime	6759.432 <prime <=11761.164	1	0.7066	0.00977	5232.2018	<.0001
puiss_fisc	5<PUISSANCE_FISCALE <=7	1	0.0713	0.0128	31.1738	<.0001
puiss_fisc	7<PUISSANCE_FISCALE <=8	1	-0.1575	0.0135	135.8563	<.0001
puiss_fisc	8<PUISSANCE_FISCALE <=9	1	-0.3696	0.0154	575.0653	<.0001
puiss_fisc	9<PUISSANCE_FISCALE	1	-0.3184	0.0146	476.8819	<.0001
age_cond	29<age_conducteur<=33	1	0.0784	0.00937	70.0195	<.0001
age_cond	33<age_conducteur<=36	1	0.0488	0.00937	27.1878	<.0001
age_cond	36<age_conducteur<=39	1	0.1254	0.00957	171.6853	<.0001

Figure 22: Extrait du tableau des estimations des paramètres

Nous remarquons par exemple que les variables : « Tranche_Prime », « Profil_risque » ainsi que « contentieux », influencent négativement la probabilité de renouvellement du contrat d'un assuré.

Les 10 variables retenues de la régression logistique sont CRM, prime, profil_risque, contentieux, difficulte, valeur_glaces, puiss_fisc, combustion, age_vehicule.

La Figure 23 fournit un extrait des rapports de cote odds.

Estimations des rapports de cotes			
Effet	Valeur estimée du point	95% Intervalle de confiance de Wald	
Tranche_prime 11761.164 <prime <=13067.966 vs prime<=2672.784	2.438	2.385	2.492
Tranche_prime 13067.966 <prime <=16898.580 vs prime<=2672.784	2.862	2.805	2.920
Tranche_prime 16898.580 <prime vs prime<=2672.784	3.221	3.149	3.295
Tranche_prime 2672.784 <prime <=3175.545 vs prime<=2672.784	1.479	1.451	1.508
Tranche_prime 3175.545 <prime <=3528.312 vs prime<=2672.784	1.689	1.652	1.728
Tranche_prime 3528.312 <prime <=3593.732 vs prime<=2672.784	2.111	2.071	2.151
Tranche_prime 3593.732 <prime <=5069.520 vs prime<=2672.784	2.130	2.083	2.178
Tranche_prime 5069.520 <prime <=6759.432 vs prime<=2672.784	2.252	2.208	2.296
Tranche_prime 6759.432 <prime <=11761.164 vs prime<=2672.784	2.027	1.989	2.066
puiss_fisc 5<PUISSANCE_FISCALE <=7 vs PUISSANCE_FISCALE<=5	1.074	1.047	1.101
puiss_fisc 7<PUISSANCE_FISCALE <=8 vs PUISSANCE_FISCALE<=5	0.854	0.832	0.877
puiss_fisc 8<PUISSANCE_FISCALE <=9 vs PUISSANCE_FISCALE<=5	0.691	0.670	0.712
puiss_fisc 9<PUISSANCE_FISCALE vs PUISSANCE_FISCALE<=5	0.727	0.707	0.748
age_cond 29<age_conducteur<=33 vs 61<age_conducteur	1.082	1.062	1.102
age_cond 33<age_conducteur<=36 vs 61<age_conducteur	1.050	1.031	1.070
age_cond 36<age_conducteur<=39 vs 61<age_conducteur	1.134	1.113	1.155
age_cond 39<age_conducteur<=46 vs 61<age_conducteur	1.161	1.143	1.180
age_cond 46<age_conducteur<=50 vs 61<age_conducteur	1.283	1.259	1.308
age_cond 50<age_conducteur<=55 vs 61<age_conducteur	1.219	1.197	1.242
age_cond 55<age_conducteur<=61 vs 61<age_conducteur	1.316	1.291	1.342
tranche_crm 100<CRM vs CRM<=90	0.531	0.511	0.552
tranche_crm 90<CRM<=100 vs CRM<=90	0.433	0.428	0.439

Figure 23: Les 22 premières lignes du tableau des estimations des rapports de cotes

Un assuré qui paye une prime de 10000 DH a 2.438 plus de chance de renouveler son contrat qu'un assuré qui paye moins de 2672 DH. De même un conducteur âgé de 30ans a 1.082 plus de chance de renouveler son contrat d'assurance automobile qu'un conducteur âgé de 62. Lorsque toute les modalités sont croisées, nous obtenons chaque profil risqué avec son coefficient correspondant.

La Figure 24 fournit les résultats des tests de concordance. Il s'agit de compter les paires concordantes, une paire d'observation étant concordante lorsque l'une vérifie cible=0, l'autre cible=1, et que la probabilité estimée que cible=1 est plus grande pour l'observation.

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	65.0	D de Somers	0.304
Pourcentage discordant	34.6	Gamma	0.305
Pourcentage lié	0.5	Tau-a	0.131
Paires	209419148510	c	0.652

Figure 24: Association des probabilités prédites et valeurs estimées

L'aire sous la courbe ROC sur l'échantillon d'apprentissage est fournie par l'indicateur « c » de Figure 24. Elle vaut donc ici « 0,652 ». La valeur de l'AUC nous montre que la capacité de prédiction n'est pas si bonne.

Le modèle régression logistique a permis de :

- Sélectionner les variables les plus discriminantes
- Regrouper les modalités
- Analyser les performances de l'AUC

L'inconvénient de ce modèle est qu'il ne prend pas en compte les interactions entre les variables explicatives.

4.4. Modèle 2 : Arbre de décision

Le modèle arbre de décision est appliqué dans cette partie pour prédire le renouvellement du contrat d'assurance.

Le module SAS/STAT ne fournit pas d'estimation d'arbre de décision, il faut faire appel au module SAS Enterprise Miner. Cause de non disponibilité de ce dernier, nous avons fait recours à IBM SPSS en s'appuyant sur l'algorithme CART pour la prédiction.

La procédure sous SPSS crée un modèle de segmentation basée sur un arbre. Elle classe les observations en groupes et estime les valeurs d'une variable (cible) dépendante à partir des valeurs de variables (prédicteur) indépendantes.

Nous obtenons la représentation graphique de l'arbre de décision ci-dessous avec le logiciel SPSS :

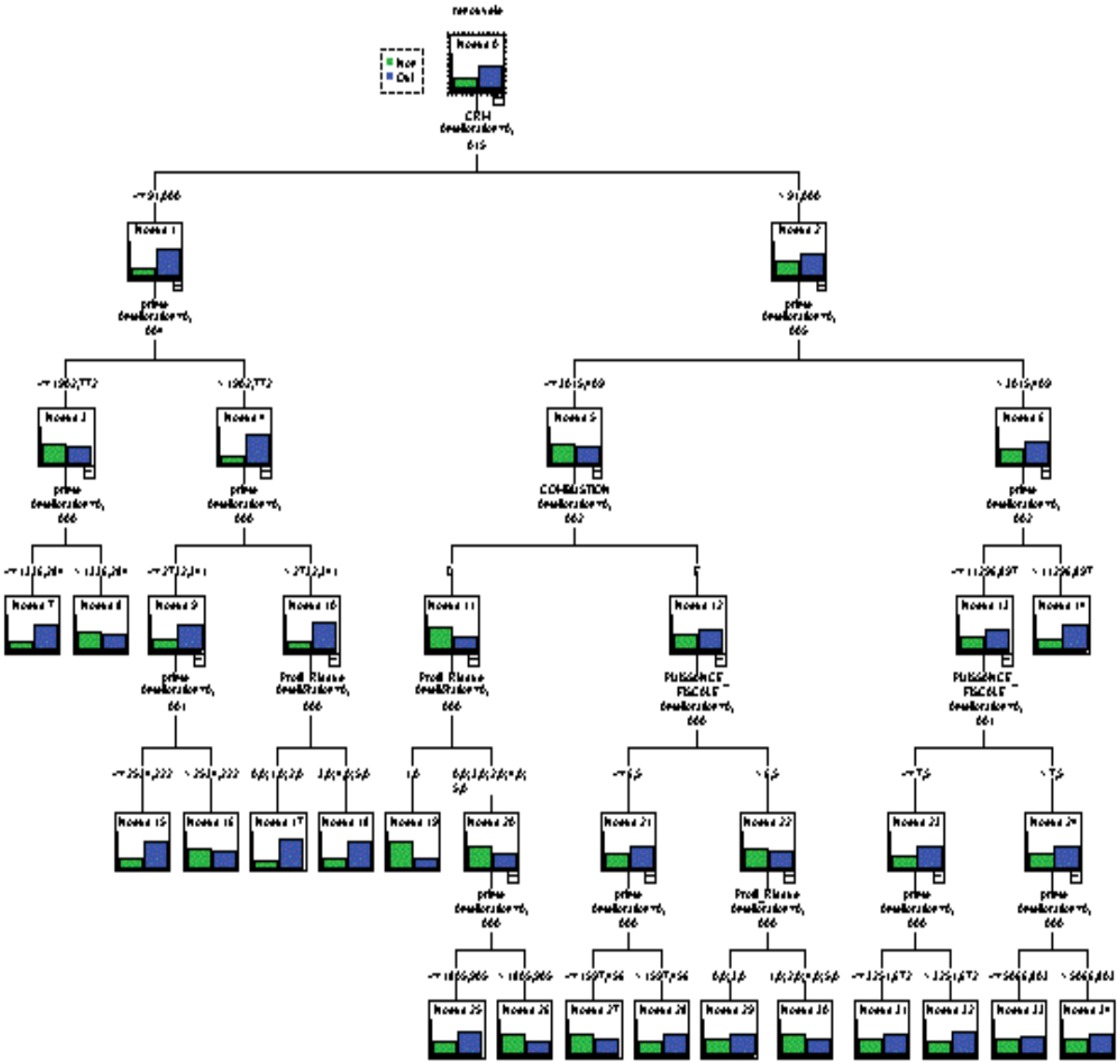


Figure 25: Représentation du modèle arbre de décision

La variable qui découpe le premier nœud est la variable CRM.

Les variables retenues de l'arbre de décision sont : CRM, profil de risque, prime et puissance fiscale.

La Figure 26 présente les résultats de faire une bonne prédiction.

Classification

Observations	Prévisions		
	Non	Oui	Pourcentage correct
Non	54292	300635	15,3%
Oui	35771	703584	95,2%
Pourcentage global	8,2%	91,8%	69,3%

Figure 26: Résultats de classification de l'arbre de décision sur la base d'apprentissage

Dans la Figure 27 produite par IBM SPSS Decision Tree, l'estimation vaut 0,307 ce qui signifie qu'un peu plus d'un client sur 3 est mal classé par l'arbre.

Risque

Estimation	Erreur standard
,307	,000

Figure 27: Coût d'un arbre de décision

La prochaine partie est consacrée à l'application des 2 modèles sur la base test. Par la suite, analyser les performances des modèles en se basant sur deux indicateurs la matrice de confusion et la courbe *ROC*.

4.5. Application sur la base test

Modèle 1

Nous appliquons le modèle 1 sur la base test.

Comme nous venons de le voir, le modèle que nous avons retenu après avoir effectué la modélisation logistique et l'arbre de décision est le modèle où interviennent les variables suivantes « garantie_ann, Tranche_CRM, Tranche_age_an, Puiss_fisc et situation_matrimoniale ».

Il s'agira maintenant d'appliquer ce modèle sur l'échantillon test afin de le valider et s'assurer de sa pertinence.

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	65.9	D de Somers	0.323
Pourcentage discordant	33.6	Gamma	0.325
Pourcentage lié	0.5	Tau-a	0.137
Paires	2377516128	c	0.662

Figure 28: Association des probabilités prédites et observées

Ces probabilités sont obtenues grâce à l'option INMODEL de la procédure LOGISTIC de SAS.
Il correspond au rapport entre le nombre total d'observations mal classées et l'effectif total dans l'échantillon.

Modèle 2

Appliqué au même fichier test, nous obtenons la matrice de confusion qui confronte les valeurs observées de la variable « *renouvele_2* » sur l'échantillon et celles prédites par le modèle, permettant ainsi de dégager le taux d'erreur qui représente l'estimation de la probabilité de mal classer.

Rappelons que la base test comporte 273851 assurés (20% de la variable initiale).
On applique notre modèle avec les 4 variables explicatives retenues du modèle 2 sur la base d'apprentissage.

Classification

	Prévisions		
Observations	0	1	Pourcentage correct
0	175508	9366	94,9%
1	75834	13143	14,8%
Pourcentage global	91,8%	8,2%	68,9%

Figure 29: Résultats de classification de l'arbre de décision sur la base test

- Le taux d'erreur du modèle 2 sur la base test est le suivant :

$$Taux\ d'erreur = \frac{9366 + 75834}{9366 + 75834 + 175508 + 13143} = 31.11\%$$

Nous avons 31.11% de chances de faire une prédiction erronée.

- La sensibilité est le vrai pourcentage d'événements prédits, c'est le taux des vrais contrats renouvelés.

$$Sensibilité = \frac{175508}{175508 + 9366} = 94.9\%$$

- La spécificité le taux des vrais contrats non renouvelés.

$$\text{spécificité} = \frac{13143}{13143 + 9366} = 57.7\%$$

- TFP = 1-Spécificité : le pourcentage de faux positifs, il s'agit de faux contrats renouvelés.

$$TFP = 1 - Spécificité = 42.3\%$$

4.6. Modèle 3 : Application sur SAS Entreprise Guide

Le but n'est pas de choisir ou comparer les modèles prédictifs, mais de construire un modèle qui prédit le mieux l'acte de renouvellement en assurance automobile.

Par ailleurs, SAS Entreprise Guide permet de construire un modèle en s'appuyant sur les principales méthodes de Data Mining.

Sas entreprise Guide nous a permis de faire notre application à travers *SAS Rapid Predictive Modeler*. Nous l'appliquons sur la base initiale qui comportait 1368133 individus.

The SAS Rapid Predictive Modeler propose les modèles Classique, Intermédiaire et Avancé. Les modèles sont de sophistication et de complexité croissantes.

- Le modèle Classique est une analyse de régression simple.

- Le modèle Intermédiaire comprend une analyse plus sophistiquée, plus l'analyse du modèle classique, et choisit le meilleur modèle.

- Le modèle Avancé comprend une analyse plus sophistiquée, plus l'analyse du modèle Classique et du modèle Intermédiaire, et choisit le meilleur modèle.

Classique

Le modèle Classique effectue successivement trois opérations de data mining.
La sélection de variables puis la transformation qui repose sur la discrétisation optimale. Cette transformation compense les valeurs de variable manquantes, et aucune imputation des valeurs manquantes n'est donc effectuée. Et finalement, la modélisation, le modèle Classique utilise un modèle de régression ascendante. Ce modèle choisit les variables une par une dans le cadre d'un processus pas à pas, qui consiste à ajouter une variable à la fois à l'équation linéaire jusqu'à ce que la contribution des variables soit insignifiante. Le modèle de régression ascendante vise à exclure les variables dépourvues d'intérêt prédictif (ou fortement corrélées avec d'autres variables de prédiction) du processus analytique.

Intermédiaire

Le modèle Intermédiaire effectue successivement sept opérations de data mining.
Sélection de variables : le modèle Intermédiaire choisit les 200 premières variables pour la modélisation.
Transformation : le modèle Intermédiaire soumet les 200 premières variables sélectionnées pour la modélisation à une transformation de puissance optimale. Ce type de transformation est un sous-ensemble de la catégorie générale de transformations de **Box-Cox**. La transformation de puissance optimale évalue un sous-ensemble de transformations de puissance exponentielle et choisit celle qui offre les meilleurs résultats pour le critère spécifié.
Imputation : le modèle Intermédiaire effectue une imputation pour remplacer les variables manquantes par les valeurs de variable moyennes.
Sélection de variables : le modèle Intermédiaire utilise les tests de critères Khi-2 et R-carré pour supprimer les variables qui ne sont pas associées à la variable à expliquer.
Combinaison de techniques de sélection de variables : le modèle Intermédiaire fusionne l'ensemble des variables sélectionnées par les tests de critères Khi-2 et R-carré.
Modélisation : le modèle Intermédiaire soumet les données d'apprentissage à trois algorithmes de modèles concurrents. **Un arbre de décision, une régression logistique et une régression pas à pas**. Dans le cas du modèle de régression logistique, les données d'apprentissage sont tout d'abord soumises à un arbre de décision qui crée une variable NODE_ID transmise au modèle de régression en tant que variable explicative. La variable NODE_ID est créée pour permettre l'élaboration de modèles d'interaction de variables.

Sélection du meilleur modèle : le modèle Intermédiaire effectue une évaluation analytique des performances de prévision ou de classification des modèles concurrents. Le modèle offrant les meilleures performances est sélectionné pour procéder à l'analyse de modélisation. Le modèle Intermédiaire pour la sélection du meilleur modèle évalue les performances non seulement des modèles intermédiaires, mais aussi des modèles classiques.
Après avoir sélectionné le meilleur modèle intermédiaire, SAS Rapid Predictive Modeler en compare ses performances de prévision avec celles du modèle classique, puis choisit le modèle optimal.

Le modèle **Avancé** effectue successivement sept opérations de data mining qui sont les suivantes :

1) Sélection de variables : le modèle Avancé choisit les 400 premières variables pour la modélisation.

2) Transformation : le modèle Avancé applique l'algorithme de transformation multiple aux 400 variables sélectionnées pour la modélisation. L'opération de transformation multiple crée plusieurs transformations de variables destinées à être utilisées dans des sélections de variables ultérieures. Les transformations multiples entraînent une augmentation du nombre de variables explicatives. Du fait de cette augmentation, SAS *Rapid Predictive Modeler* sélectionne les 400 meilleures variables explicatives parmi les résultats générés par l'algorithme de transformation multiple.

3) Imputation : le modèle Avancé effectue une imputation pour remplacer les variables manquantes par les valeurs de variable moyennes. Cette opération crée également des variables indicatrices qui permettent à l'utilisateur d'identifier les observations contenant des valeurs de variable imputées.

4) Sélection de variables : le modèle Avancé utilise les tests de critères Khi-2 et r-carré pour supprimer les variables qui ne sont pas associées à la variable à expliquer.

5) Combinaison de techniques de sélection de variables : le modèle Avancé fusionne l'ensemble des variables sélectionnées par les tests de critères Khi-2 et R-carré.

6) Modélisation : le modèle Avancé soumet les données d'apprentissage à quatre algorithmes de modèles concurrents. Les modèles sont : un arbre de décision, un modèle de réseau neuronal, un modèle de régression descendante et un modèle d'ensemble. Le modèle de réseau neuronal effectue des recherches restreintes afin de détecter un réseau ascendant optimal. La régression descendante est un modèle de régression linéaire qui élimine des variables en les retirant une par une à la fois jusqu'à ce que les scores du test R-carré chutent de manière significative. Le modèle d'ensemble crée de nouveaux modèles en combinant les probabilités a posteriori (pour les variables qualitatives à expliquer) ou les valeurs prédites (pour les variables continues à expliquer) de plusieurs modèles d'entrée prédécesseurs.
   Le nouveau modèle d'ensemble est alors utilisé pour procéder au scoring de nouvelles données. Le modèle d'ensemble utilisé dans le modèle avancé est créé à partir du résultat du modèle classique, du meilleur modèle du modèle intermédiaire, et du meilleur modèle du modèle avancé.

7) Sélection du meilleur modèle : le modèle avancé effectue une évaluation analytique des performances de prévision ou de classification des modèles concurrents d'arbre de décision, neuronaux et de régression. Le modèle qui offre les meilleures performances de prévision ou de classification est alors utilisé en entrée, avec les meilleurs modèles classique et intermédiaire, pour former un modèle d'ensemble. Ensuite, les nouveaux modèles d'ensemble, d'arbre de décision, neuronaux et de régression descendante sont soumis à une analyse comparative afin de déterminer le modèle optimal parmi les échantillons des meilleurs modèles classiques, intermédiaires et avancés.

Une fois que le SAS *Rapid Predictive Modeler* a sélectionné le meilleur modèle, il exécute et compare les performances de prévision du modèle Avancé avec les meilleurs modèles Intermédiaire et Classique, puis choisit le plus performant.

Application

Rappelons que notre base initiale contient 1368133 assurés et 164 variables explicatives. Sur SAS Entreprise Guide, nous analysons les résultats d'application d'un modèle prédictif avec l'ensemble de variables explicatives. On choisit le modèle avancé de *SAS Rapid Predictive Modeler.*

Propriétés	Valeur
Source de données d'entrée	SASUSER.BASE_INITIALE
Variable à expliquer	renouvele
Niveau d'événement	OUI
Observations	1368133
Variables originales	164
Variables sélectionnées	33

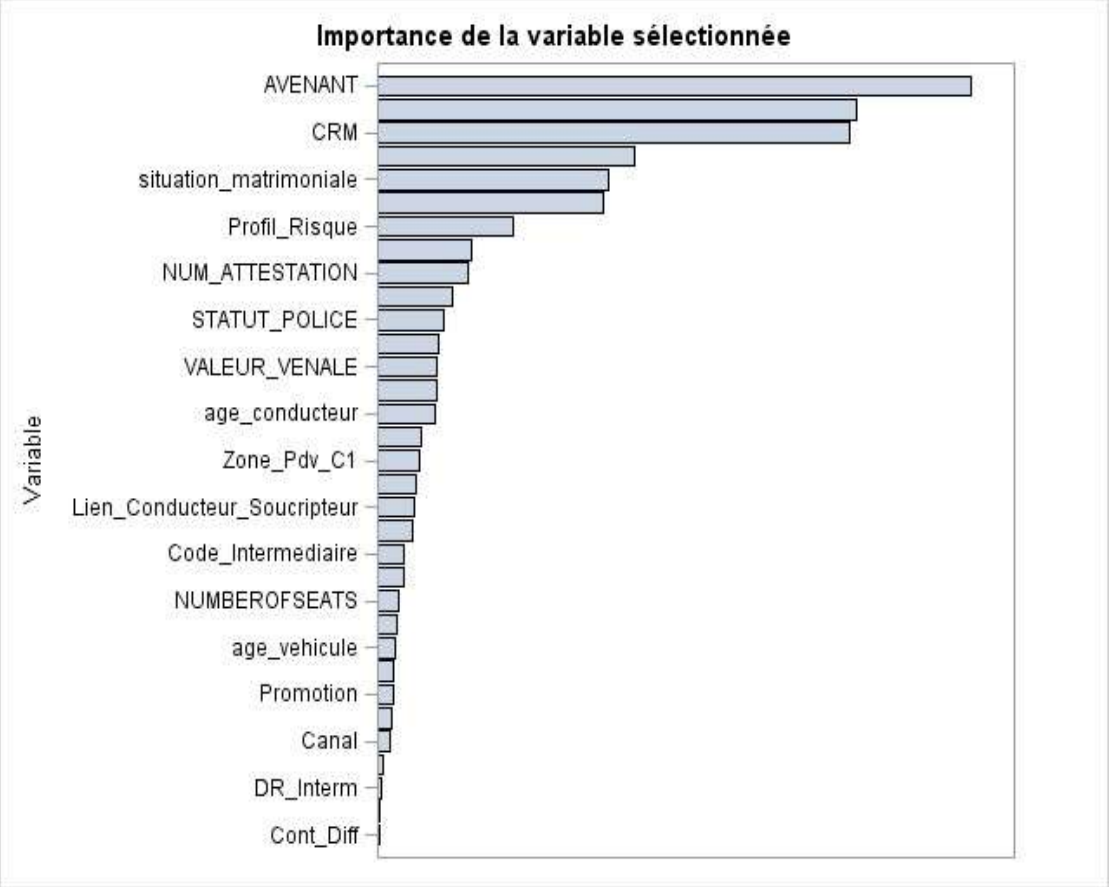


Figure 30: Classement des variables sélectionnées

Le modèle retient 33 variables explicatives des 165. La variable « avenant » et « CRM » sont les variables qui influencent plus l'acte de renouvellement.

Cible de la matrice de classification =renouvele

	Rôle des données			
	TRAIN		VALIDATE	
	Predicted		Predicted	
	NON	OUI	NON	OUI
A expliquer				
NON	30.16	69.84	26.48	73.52

OUI	8.89	91.11	10.85	89.15

Figure 31: Matrice de confusion du modèle 3

La courbe ROC est représentée après l'établissement de la table « *Receiver Operating Characteristict* » où sont représentés les valeurs de la sensibilité ainsi que celles de (1-la spécificité).

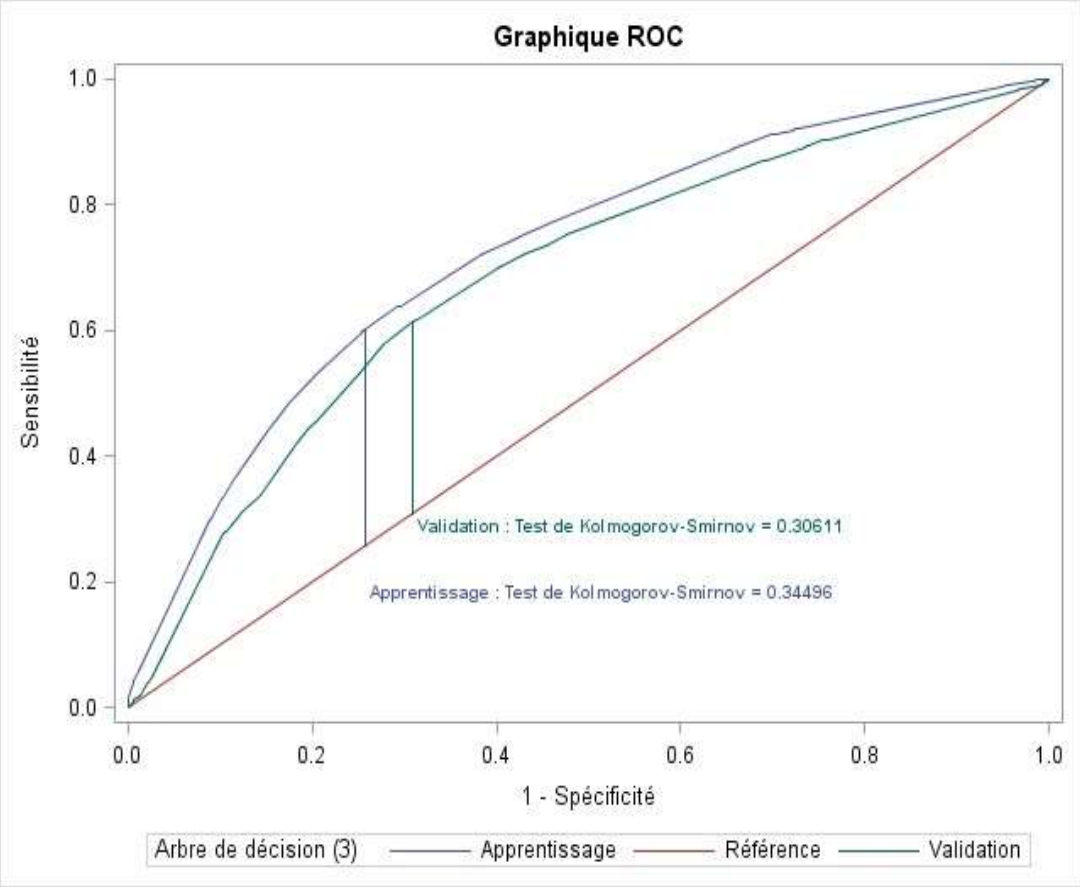


Figure 32: Courbe ROC du modèle 3

Plus la courbe ROC s'approche du coin supérieur gauche du carré de la Figure 32 et plus elle s'éloigne de la première bissectrice, meilleur est le modèle, car elle permet de capturer le plus possible de vrais évènements (vrais *churners*) avec le moins possible de faux évènements (faux *churners*). Le modèle est accepté.

Sélection de modèle basée sur Valid: Average Profit for renouvele

Modèle sélectionné	Noeud du modèle	Description du modèle	Libellé cible	Train: Akaike's Information Criterion
Y	Tree3	Arbre de décision (3)		.
	Reg5	Régression pas à pas		27740.12
	Ensmbl	Ensemble_Champion		.
	Reg2	Régression ascendante		34170.43

Apprentissage : Levier (Lift)	Validation : Levier (Lift)
1.27967	1.27525
1.28262	1.28123
1.29833	1.31368
1.00000	1.00000

Figure 33: Résultats des modèles sur SAS Rapid Predictiv Model

Statistique	Train	Validate
Somme des fréquences	23999.0000	16001.0000
Taux de mauvaise classification	0.2867	0.3119
Erreur absolue maximale	0.9655	1.0000
Erreurs de la somme des carrés	9056.7712	6431.1864
Erreur quadratique approchée (ASE : Aver	0.1887	0.2010
Racine carrée de l'erreur quadratique mo	0.4344	0.4483
Index Roc	0.7240	0.6850
Coefficient de Gini	0.4480	0.3690
Test de Kolmogorov-Smirnov	0.3450	0.3060
Seuil de la probabilité de Kolmogorov-S	0.6800	.
Levier (Lift) à 10%	1.2797	1.2753
Levier (Lift) cumulé à 10%	1.3144	1.2320
Réponse capturée à 10%	6.3986	6.3759
Réponse capturée cumulée % à 10%	13.1442	12.3266

Figure 34: Statistiques du modèle 3

- Taux d'erreur du modèle

Le taux de faire une mauvaise prédiction est de 28.67% sur la base d'apprentissage et il est de 31.19% sur la base test.

- ROC

72.4% sur la base d'apprentissage et 68.5% sur la base test. Notre modèle semble accepté pour prédire l'acte de renouvellement d'un contrat d'assurance automobile.

III. Conclusion

Les techniques datamining nous ont permis d'avoir une idée sur les clients qui restent dans le portefeuille ainsi que ceux qui le quittent. Le phénomène que nous avons cherché à modéliser est l'acte de renouvellement d'un contrat d'assurance automobile.

En premier temps, le nombre important des variables explicatives qui est de 165 variables, nous a poussé à réduire le périmètre de ces variables en s'appuyant sur les méthodes de sélection des variables, donc de garder celle qui influencent le plus le choix d'un client pour renouveler son contrat ou non.
L'échantillonnage est l'étape fondamentale pour appliquer un modèle prédictif. Nous avons opté pour la méthode classique appelée « *Cross-Validation* ».

Chaque modèle prédictif a sa particularité, nous avons pensé à combiner quelques méthodes pour améliorer la performance du modèle et mieux analyser les résultats, le modèle 2 est le modèle combinant les 3 méthodes.

Le modèle 3 obtenu sous SAS Entreprise Guide avec l'outil *SAS Rapid Predictive Model* semble le modèle qui prédit le mieux l'acte de renouvellement sur la base d'apprentissage. Le taux de faire une mauvaise prédiction est de 28.67% sur la base d'apprentissage et de 31.19% sur la base test.

Cependant, ce modèle a été appliqué sur 165 variables explicatives, 33 variables ont été sélectionnées. Sachant que notre base initiale compte un nombre important de données manquantes, *SAS Rapid Predictive Model* remplace chaque valeur manquante par la moyenne des valeurs de la variable explicative. Cette technique d'imputation n'est pas autant fiable et peut biaiser le modèle.

Le Modèle 3 est accepté en terme de meilleur prédiction, or en terme de stabilité le modèle 2 est meilleur comme le taux d'erreur sur la base d'apprentissage est proche du taux d'erreur sur la base test.

Néanmoins, pour améliorer le modèle une attention particulière doit porter sur les variables qui comportent des données manquantes. En effet, des méthodes de type "bootstrap" peuvent être utilisées pour combler ces données manquantes. Ce type de méthodes peut convenir pour améliorer le taux d'erreur des modèles utilisés dans ce mémoire. En effet, comme nous avons pu le voir avec le modèle 3 qui remplace les données manquantes par une simple moyenne, la stabilité des résultats et des performances de ce dernier, entre la base test et la base d'apprentissage, ont été relativement bien augmentés. Il est donc souhaitable d'étudier ces variables et utiliser leur comportement statistique afin de "combler" leurs données manquantes.

Bibliographie

A, TREILHOU. 2000. *ELABORATION D'UN MODELE DE RESILIATION DES CONTRATS AUTO.* 2000.

Arthur, CHARPENTIER. 2015-2016. *ACTUARIAT DE L'ASSURANCE NON VIE (UQAM ET UNIVERSITE DE RENNES) #2.* [http://freakonometrics.hypotheses.org] 2015-2016.

Assurance, Saham. http://www.sahamassurance.com/. [En ligne]

FMSR. 2006. *Fédération Marocaine des sociétés d'assurance et de réassurace.* [En ligne] juin 2006. https://www.fmsar.org.ma/docs/manuel-d-aide-a-la-consultation-du-CRM.pdf.

GONNET, GUILLAUME. 2010. *ETUDE DE LA TARIFICATION ET DE LA SEGMENTATION EN ASSURANCE AUTOMOBILE.* 2010.

Guide, SAS Entreprise. Aide de Sas Entreprise Guide.

http://www.financialafrik.com/. *Financiel Afrik.* [En ligne]

LEJEUNE, FRANÇOIS-XAVIER. *INTRODUCTION AU LOGICIEL SAS.* s.l. : INSTITUT DE STATISTIQUE DE L'UNIVERSITE PIERRE ET MARIE CURIE.

—. INTRODUCTION AU LOGICIEL SAS. s.l. : INSTITUT DE STATISTIQUE DE L'UNIVERSITE PIERRE ET MARIE CURIE.

Maroc, SAHAM assurance. 2013. *Rapport annuel de Saham assurance.* 2013.

MARYLENE, CUBBER. 2011. *RENTABILITE ET TARIFICATION SOUS SOLVABILITE II.* 2011.

MENARD, CELINE BOUQUET ET PHILIPPE. 2011. *Assurance automobile-Optimisation des ressources à l'échéance .* 2011.

NAKACHE J.P., CONFAIS J. 2005. *Approche pragmatique de la classification.* s.l. : TECHNIP, 2005.

Rakotomalala, Ricco. 2015. *Pratique de la Régression Logistique.* s.l. : Univerrsité Lumière Lyon 2, 2015.

RICCO, RAKOTOMOLALA. PRATIQUE DE LA REGRESSION LOGISTIQUE, REGRESSION. [En ligne]

THOMAS, BOUCHE. *MODELE DE PROPENSION DES ASSURES PAR RAPPORT AUX RISQUES DE SINISTRES CORPORELS GRAVES EN ASSURANCE AUTOMOBILE.*

Tufféry, Stéphane. 2005. *Data mining et statistique décisionnelle.* s.l. : Edition TECHNIP, 2005.

Annexes

Annexe 1 : Variables spécifiques du contrat

Nom	Libellé
AVENANT	Avenant
CONTRAT_PARRAIN_FILLEUL	Contrat ayant l'option parrain-filleul
DATE_EXIGIBILITE	Date d'exigibilité
DATE_RESILIATION	Date de résiliation
DEBUT	Début du contrat
FIN	Fin du contrat
FORMULE	Formule du contrat
Gamme	Gamme de l'assurance
ID_PRODUIT	Identifiant du produit
NATURE_DU_CONTRAT	Nature du contrat
Prime	Prime du contrat
renouvele	Renouvellement du contrat
STATUT_POLICE	Statut de la police

Table 8: Variables liées au contrat

Nom	Libellé
CRM	Coefficient de réduction majoration
CSP	Catégorie socio professionnelle
Date_Naissance	Date de naissance du souscripteur
Lien_Conducteur_Soucripteur	Lien entre le conducteur et le souscripteur
NOMBRE_DE_CONDUCTEUR	
Nombre_enfants	Nombre des enfants
Sexe	Sexe du souscripteur
situation_matrimoniale	Situation matrimoniale

Table 9: Les données liées au « conducteur »

La variable "Lien_conducteur_soucripteur » prend 7 modalités :

Lien_Conducteur_Soucripteur	conjoint	Conjointe	employé	enfant	Parent	Souscripteur	Autre

Annexe 2 : Les données « Garanties »

Dans cette section, nous avons les variables ayant rapport avec les garanties fournies dans les différents contrats

Notre base contient 30 variables de différentes garanties du produit d'assurance automobile

Nom	Libellé
Actes_de_vandalismes_YN	Actes de vandalismes
Amenagements_professionnels_YN	Aménagements professionnels
Bris_des_retroviseurs_YN	Bris des rétroviseurs
Bris_de_glaces_YN	Bris de glaces
Collision_YN	Collision
Collision_Etendue_YN	Collision étendue
Defense_Recours_YN	Défense du recours
Dommages_au_vehicule_limite_YN	Dommages au véhicule limités
Dommages_au_vehicule_YN	Dommages au véhicule
Dommages_tous_risque_YN	Dommages tous risques
Evenements_climatiques_YN	Evènements climatiques
GARANTIE_ANNEXES	Garanties annexes
Incendie_1er_evenement_YN	1er évènement d'incendie
Inondation_YN	Inondation
Marchandises_transportees_YN	Marchandises transportées
MRH_YN	Garantie multirisque habitation
MRP_YN	Garantie multirisque professionnelle
Personnes_transportees_YN	Personnes transportées
Perte_Financiere_YN	Perte financière
Pro_du_Cond_et_des_Pass_YN	Protection des conducteurs et des passagers
Protection_Conducteur_YN	Protection du conducteur
Protection_Juridique_YN	Protection juridique
Rachat_de_vetuste_YN	Rachat de vétusté
Valeur_Majoree_YN	Valeur majorée
Vol_1er_evenement_YN	1er évènement de vol

Vol_de_la_roue_de_secours_YN	Vol de la roue de secours
Vol_des_retroviseurs_YN	Vol des rétroviseurs
Vol_et_Incendie_1er_Evenement_YN	1er évènement de vol ou d'incendie
Vol_materiel_audio_video_YN	Vol du matériel audio ou vidéo
Vol_YN	Vol

Table 10: Liste des données garanties

Annexe 3 : Analyse statistique de quelques variables quantitatives

Variable	N	Moyenne	Ecart-type	Minimum	Maximum
CRM	1368133	96.4664226	5.93	90%	210%
Age_conducteur	1368133	44.5284435	12.44	10	113
VALEUR_NEUVE	1368133	45892.32	8562426.96	0	9999999999
Ind_BDG	1368133	0.1545456	0.3614711	0	1

Table 11: Analyse statistique de 4 variables quantitatives

On peut remarquer que la variable « age_conducteur » comporte des valeurs erronées, sur le tableau d'analyse statistique le minimum de cette variable est 10. Cependant, l'âge minimum pour avoir un permis de conduite au Maroc est de 18 ans.

Annexe 4 : Arbre de décision « Discrétisation par l'algorithme CHAID »

Ci-dessous la représentation graphique de la classification des variables « age_conducteur » et « prime » par les arbres de décision à l'aide d'algorithme CHAID.

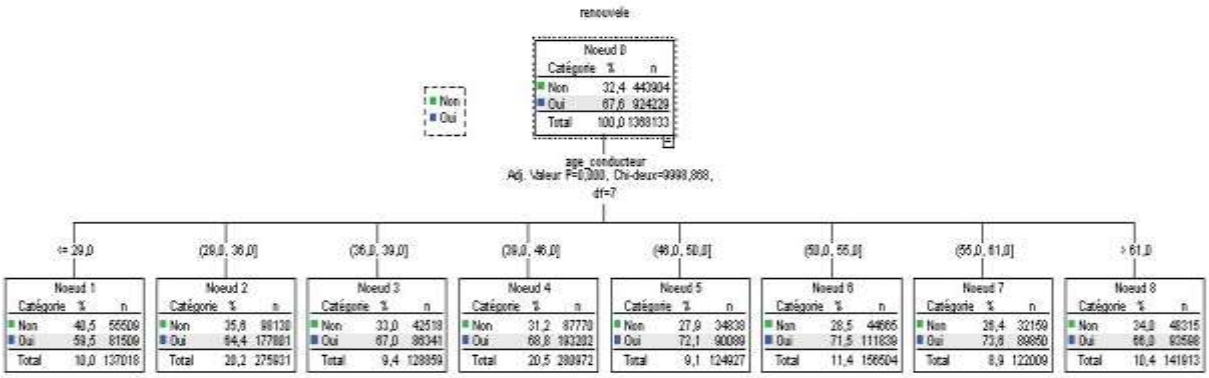


Figure 35: Discrétisation de la variable "age_conducteur" par arbre de décision

La variable « age_conducteur » est découpée en 8 nœuds. Nous remarquons que la tranche d'âge [56,61] est la tranche ou le conducteur a tendance à renouveler son contrat, tandis que le conducteur âgé de moins de 29ans est le plus qui a tendance à ne pas renouveler son contrat.

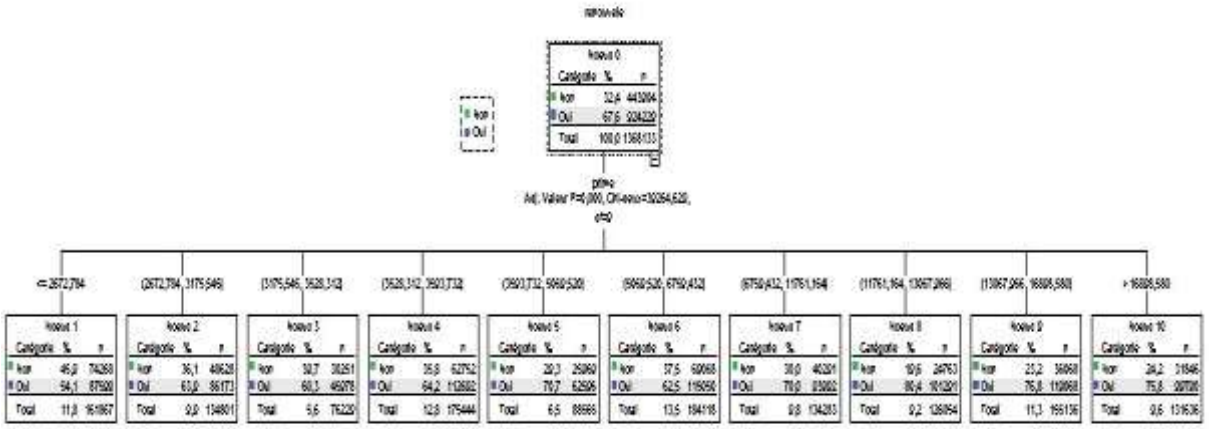


Figure 36: Discrétisation de la variable "prime" par arbre de décision

La variable « prime » est découpée en 10 nœuds. Les assurés qui ne renouvellent pas leurs contrats sont les assurés qui payent moins leurs primes. Selon la Figure 36, 45% de clients ne renouvellent pas leurs contrats quand ils payent un montant de prime inférieur à 2672 DH.

Annexe 5: Résultats de la modélisation

- **L'analyse discriminante** :

Le Système SAS

The DISCRIM Procedure

Generalized Squared Distance to renouvele_2

De renouvele_2	0	1
0	0.76044	2.50638
1	0.96463	2.30219

Fonction discriminante linéaire pour renouvele_2

Variable	0	1
Constant	-17788	-17794
CRM	3.55170	3.60615
MONTANT_ACCESSOIRES	3.93596	3.92667
POIDS_EN_CHARGE	0.0000668	0.0000783
Profil_Risque	-6.31584	-6.32749
PUISSANCE_FISCALE	0.39674	0.46925
Valeur_Conventionnelle	-1.2258E-6	-1.0778E-6
VALEUR_DES_GLACES	-0.00246	-0.00246
VALEUR_NEUVE	1.22733E-6	1.07974E-6
VALEUR_VENALE	-0.0000106	-0.0000104
Code_PDV	0.0000627	0.0000619
Ind_BDG	4.76805	4.74316
Ind_Collision	0.47194	0.38903
Ind_Tierce	35187	35188
Ind_Incendie	-8.60162	-8.19153
Ind_Perte_Fin	43.76169	45.31283
Ind_Vol	10.21048	9.99700
Ind_Pro_Cond_Pass	-43.15420	-43.20987
Ind_Coll_Etendue	5.12120	5.12382
Ind_Pro_Juri	2.35495	2.27870
Ind_Vol_Inc_1er	-16.48796	-16.29509
Ind_Inondation	27.78218	28.46202
Ind_DTR	35100	35101
Ind_Tierce_Limite	-3.74031	-3.87554
Indic_Dommage	35217	35217
prime	0.0009349	0.0008907
age_permis	-0.16536	-0.17079
age_conducteur	0.48701	0.49296
age_vehicule	0.00582	-0.0009550

- **Modèle logistique**

Les résultats de l'estimation descendante du modèle logistique.

Procédure d'élimination descendante

Informations sur les niveaux de classe

Classe	Valeur	Variables d'expérience				
Contentieux	Non	1				
	Oui	0				
Difficulte	Non	1				
	Oui	0				
Cont_Diff	Non	1				
	Oui	0				
Ind_BDG	0	1				
	1	0				
age_veh	13<age_vehicule	1	0	0	0	0 0 0 0
	17<age_vehicule	0	1	0	0	0 0 0 0
	21<age_vehicule	0	0	1	0	0 0 0 0
	25<age_vehicule	0	0	0	1	0 0 0 0
	29<age_vehicule	0	0	0	0	1 0 0 0
	2<age_vehicule<	0	0	0	0	0 1 0 0
	5<age_vehicule<	0	0	0	0	0 0 1 0
	7<age_vehicule<	0	0	0	0	0 0 0 1
	age_vehicule<=2	0	0	0	0	0 0 0 0
COMBUSTION	D	1				
	E	0				
Ind_Collision	0	1				
	1	0				
tranche_crm	100<CRM	1	0			
	90<CRM<=100	0	1			
	CRM<=90	0	0			
puiss_fisc	5<PUISSANCE_FISCALE <=7	1	0	0	0	
	7<PUISSANCE_FISCALE <=8	0	1	0	0	
	8<PUISSANCE_FISCALE <=9	0	0	1	0	
	9<PUISSANCE_FISCALE	0	0	0	1	
	PUISSANCE_FISCALE<=5	0	0	0	0	
Indic_Dommage	0	1				
	1	0				
Ind_Vol	0	1				
	1	0				
Ind_Incendie	0	1				
	1	0				
Profil_Risque	0	1	0	0	0	0
	1	0	1	0	0	0
	2	0	0	1	0	0
	3	0	0	0	1	0
	4	0	0	0	0	1
	5	0	0	0	0	0
age_cond	29<age_conducteur<=33	1	0	0	0	0 0 0
	33<age_conducteur<=36	0	1	0	0	0 0 0
	36<age_conducteur<=39	0	0	1	0	0 0 0
	39<age_conducteur<=46	0	0	0	1	0 0 0
	46<age_conducteur<=50	0	0	0	0	1 0 0
	50<age_conducteur<=55	0	0	0	0	0 1 0
	55<age_conducteur<=61	0	0	0	0	0 0 1
	61<age_conducteur	0	0	0	0	0 0 0
situation_matrimoniale	Célibataire	1	0	0	0	0
	Divorcé	0	1	0	0	0
	Divorcé(e)	0	0	1	0	0
	Marifé(e)	0	0	0	1	0
	Marié(e)	0	0	0	0	1
	Veuf(ve)	0	0	0	0	0
Tranche_prime	11761.164 <prime <=13067.966	1	0	0	0	0 0 0 0 0
	13067.966 <prime <=16898.580	0	1	0	0	0 0 0 0 0
	16898.580 <prime	0	0	1	0	0 0 0 0 0
	2672.784 <prime <=3175.545	0	0	0	1	0 0 0 0 0
	3175.545 <prime <=3528.312	0	0	0	0	1 0 0 0 0
	3528.312 <prime <=3593.732	0	0	0	0	0 1 0 0 0
	3593.732 <prime <=5069.520	0	0	0	0	0 0 1 0 0
	5069.520 <prime <=6759.432	0	0	0	0	0 0 0 1 0
	6759.432 <prime <=11761.164	0	0	0	0	0 0 0 0 1
	prime<=2672.784	0	0	0	0	0 0 0 0 0

Effet	Estimations des rapports de cotes Valeur estimée du point	95% Intervalle de confiance de Wald	
Contentieux Non vs Oui	1.560	1.537	1.583
Difficulte Non vs Oui	0.589	0.572	0.606
Cont_Diff Non vs Oui	1.990	1.936	2.045
Ind_BDG 0 vs 1	0.922	0.899	0.945
age_veh 13<age_vehicule vs age_vehicule<=2	0.984	0.958	1.010
age_veh 17<age_vehicule vs age_vehicule<=2	1.045	1.018	1.073
age_veh 21<age_vehicule vs age_vehicule<=2	1.072	1.044	1.101
age_veh 25<age_vehicule vs age_vehicule<=2	1.038	1.011	1.066
age_veh 29<age_vehicule vs age_vehicule<=2	0.997	0.970	1.026
age_veh 2<age_vehicule< vs age_vehicule<=2	0.940	0.918	0.963
age_veh 5<age_vehicule< vs age_vehicule<=2	0.972	0.947	0.997
age_veh 7<age_vehicule< vs age_vehicule<=2	0.925	0.903	0.947
COMBUSTION D vs E	0.916	0.903	0.929
tranche_crm 100<CRM vs CRM<=90	0.587	0.561	0.614
tranche_crm 90<CRM<=100 vs CRM<=90	0.464	0.457	0.471
puiss_fisc 5<PUISSANCE_FISCALE <=7 vs PUISSANCE_FISCALE<=5	1.110	1.077	1.144
puiss_fisc 7<PUISSANCE_FISCALE <=8 vs PUISSANCE_FISCALE<=5	0.943	0.914	0.974
puiss_fisc 8<PUISSANCE_FISCALE <=9 vs PUISSANCE_FISCALE<=5	0.779	0.751	0.808
puiss_fisc 9<PUISSANCE_FISCALE vs PUISSANCE_FISCALE<=5	0.817	0.789	0.845
Indic_Dommage 0 vs 1	0.959	0.933	0.984
Ind_Vol 0 vs 1	0.871	0.846	0.896
Ind_Incendie 0 vs 1	1.256	1.222	1.292
Profil_Risque 0 vs 5	1.311	1.247	1.377
Profil_Risque 1 vs 5	1.096	1.047	1.147
Profil_Risque 2 vs 5	1.189	1.137	1.244
Profil_Risque 3 vs 5	1.158	1.106	1.212
Profil_Risque 4 vs 5	1.154	1.102	1.208
age_cond 29<age_conducteur<=33 vs 61<age_conducteur	1.042	1.019	1.066
age_cond 33<age_conducteur<=36 vs 61<age_conducteur	0.998	0.976	1.021
age_cond 36<age_conducteur<=39 vs 61<age_conducteur	1.102	1.077	1.128
age_cond 39<age_conducteur<=46 vs 61<age_conducteur	1.126	1.105	1.148
age_cond 46<age_conducteur<=50 vs 61<age_conducteur	1.220	1.193	1.248
age_cond 50<age_conducteur<=55 vs 61<age_conducteur	1.177	1.152	1.202
age_cond 55<age_conducteur<=61 vs 61<age_conducteur	1.245	1.217	1.274
situation_matrimonia Célibataire vs Veuf(ve)	0.805	0.692	0.938
situation_matrimonia Divorcé vs Veuf(ve)	0.832	0.549	1.260
situation_matrimonia Divorcé(e) vs Veuf(ve)	1.004	0.666	1.513
situation_matrimonia Marifé(e) vs Veuf(ve)	0.961	0.826	1.119
situation_matrimonia Marié(e) vs Veuf(ve)	0.975	0.837	1.135
Tranche_prime 11761.164 <prime <=13067.966 vs prime<=2672.784	1.708	1.663	1.755
Tranche_prime 13067.966 <prime <=16898.580 vs prime<=2672.784	1.906	1.857	1.957
Tranche_prime 16898.580 <prime vs prime<=2672.784	2.060	1.995	2.126
Tranche_prime 2672.784 <prime <=3175.545 vs prime<=2672.784	1.425	1.388	1.463
Tranche_prime 3175.545 <prime <=3528.312 vs prime<=2672.784	1.493	1.449	1.537
Tranche_prime 3528.312 <prime <=3593.732 vs prime<=2672.784	1.832	1.786	1.879
Tranche_prime 3593.732 <prime <=5069.520 vs prime<=2672.784	1.840	1.787	1.895
Tranche_prime 5069.520 <prime <=6759.432 vs prime<=2672.784	1.754	1.710	1.799
Tranche_prime 6759.432 <prime <=11761.164 vs prime<=2672.784	1.483	1.447	1.519

Figure 37: rapport de cote (odds ratio) de la régression logistique