RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université M'hamed Bougara de Boumerdés

Faculté des Sciences

Département des Mathématiques

Mémoire présenté ,
Par
SEIDI Rania

Pour l'obtention du diplôme de Master en Mathémathiques appliquées

Option: *Mathématiques Financières*

Sujet :

La tarification en assurance automobile

Soutenue publiquement, devant le jury composé de :

Mr	BENAMARA OUALID	Président
Mr	MAHIEDDINE ZITOUNI	Encadreur
Mme	MEDDAHI SAMIA	Examinatrice

Année universitaire:2022/2023

Résumé

L'assurance automobile est un secteur stratégique dans le domaine de l'assurance, où la responsabilité civile est une couverture obligatoire. De nombreuses études théoriques et empiriques se sont intéressées à la tarification de cette assurance, car elle revêt une grande importance en déterminant la prime que l'assuré doit payer.

Notre étude a pour objectif de déterminer les facteurs qui influencent le calcul de la prime de responsabilité civile en assurance automobile en Algérie, en utilisant une modélisation GLM et CART, implémentés à l'aide du logiciel Python (Jupyter Notebook), pour identifier les facteurs qui influencent le calcul de la prime de responsabilité civile en assurance automobile. Le modèle GLM, largement utilisé dans la tarification automobile, nous permet d'évaluer l'impact des variables sur les prédictions. Ensuite, nous examinons le modèle CART, un modèle d'arbre de régression réputé pour sa facilité de compréhension. Ces modèles nous offre une meilleure tarification.

L'analyse des éléments déterminants de la prime révèle que le système de tarification actuel en Algérie est insuffisant, et il est crucial d'intégrer de nouveaux facteurs afin de proposer une tarification plus adaptée à chaque assuré.

mots clés : Assurance automobile,Tarification,Prime,Modélisation GLM,modélisation CART.

Abstract

Automobile insurance is a strategic sector in the insurance industry, where third-party liability coverage is mandatory. Numerous theoretical and empirical studies have focused on the pricing of this insurance, as it plays a significant role in determining the premium that policyholders have to pay.

Our study aims to identify the factors that influence the calculation of the third-party liability premium in automobile insurance in Algeria, using GLM and CART

modeling implemented with Python software (Jupyter Notebook). The GLM model, widely used in automobile pricing, allows us to assess the impact of variables on predictions. Additionally, we explore the CART model, a regression tree model known for its ease of interpretation. These models provide a better pricing approach.

The analysis of the key determinants of the premium reveals that the current pricing system in Algeria is inadequate, and it is crucial to integrate new factors to propose more tailored pricing for each policyholder.

Keywords Automobile insurance, Pricing, Premium, GLM modeling, CART modeling.

Remerciements

En premier lieu, nous tenons à remercier dieu le tout puissant de nous avoir donné l'opportunité d'étudier, la volonté, le courage, la force et la patience pour surmonter toutes les difficultés rencontrées durant cette année.
Nous remercions également notre promoteur Mr MAHIEDDINE ZITOUNI qui nous a permis de bénéficier de son encadrement, les conseils qu'il nous a prodigués, la patience, la confiance qu'il nous a témoignée ont été déterminants dans la réalisation de ce mémoire.
Nous tenons également à remercier les membres de jury pour avoir accepté d'évaluer ce modeste travail.
Nos vifs remerciements à notre encadreurs de stage Mme NOUIOUA Achouak et Mr MECHAREK Abdelkader, qui nous ont suivis durant la durée du stage pratique à la société Algérienne d'assurance, pour leur soutien du début jusqu'à la fin et leur aide durant notre stage.
Nos vifs remerciements pour l'ensemble du personnel de la compagnie Algérienne des assurances de la wilaya d'ALGER pour leurs accueils et leurs disponibilités
Nous remercions aussi toutes les personnes qui ont bien voulu nous accorder un peu de leur temps et leurs connaissances.

Dédicace

Merci mon dieu de m'avoir donné le courage d'aller jusqu'au bout.
A ceux qui se sont donné la peine, leurs encouragements et leurs sacrifices pour me voir réussir dans la vie,
Au meilleur des pères ce brave homme qui a veillé à mon éducation, sacrifié sa vie pour ma réussite, je ne vous remercierai jamais assez.
A ma très chère maman qui m'a comblée avec la tendresse et affection tout au long de mon parcours, qui n'a cessé de me soutenir et de m'encourager durant toutes les années de mes études.
A mes frères Zinou et Rayan, merci d'être dans ma vie et de me soutenir.
À mes oncles, tantes, cousins et cousines ; en témoignage de mon amour, de mon Profond respect et de ma reconnaissance.
À tous mes amis pour leur soutien moral et sympathie. À tous ceux qui ont participé à ma formation.

Table des matières

Table des figures

Liste des tableaux

Introduction générale :

L'assurance générale en Algérie joue un rôle crucial dans la protection des individus, des entreprises et des biens contre les risques financiers imprévus. Ce secteur dynamique offre une gamme de produits et de services conçus pour couvrir diverses formes de risques, contribuant ainsi à la stabilité économique du pays, L'industrie de l'assurance générale en Algérie est réglementée par l'Autorité de Contrôle des Assurances et de la Prévoyance Sociale (ACAPS), qui veille à l'application des normes et des règles en matière d'assurance.

Cette réglementation vise à protéger les intérêts des assurés et à garantir l'intégrité du secteur.Les compagnies d'assurance générale en Algérie proposent une variété de produits adaptés aux besoins et aux exigences des particuliers et des entreprises. Ces produits incluent des polices d'assurance automobile, incendie, vol, responsabilité civile, dommages aux biens, responsabilité professionnelle, ainsi que des assurances voyage et santé, entre autres.L'assurance automobile est obligatoire en Algérie, offrant une protection contre les dommages causés aux véhicules, les blessures corporelles et les responsabilités légales en cas d'accidents de la route. De plus, l'assurance incendie et vol protège les biens immobiliers et le contenu contre les pertes dues à des incidents tels que les incendies, les explosions ou les vols.Pour les entreprises, l'assurance générale offre une couverture contre divers risques, tels que les dommages matériels, les pertes financières, la responsabilité civile professionnelle et les perturbations d'activité. Ces polices d'assurance aident les entreprises à se protéger contre les imprévus qui pourraient compromettre leur fonctionnement et leur rentabilité.

La tarification en assurance est un aspect essentiel à considérer, mais également

complexe. Les assureurs doivent évaluer divers facteurs tels que l'âge, l'historique des réclamations, le type de véhicule ou de propriété, la localisation géographique, etc., pour déterminer le montant de prime adéquat pour chaque assuré. Cependant, cette évaluation peut présenter des défis et des disparités de tarification. La tarification en assurance automobile repose également sur des statistiques et des données actuarielles, permettant aux assureurs d'estimer les coûts probables des sinistres et de fixer des primes qui couvrent ces coûts tout en maintenant leur rentabilité.

Il est important de noter que la tarification en assurance automobile est réglementée dans de nombreux pays pour éviter toute discrimination ou tarification excessive. Les assureurs doivent se conformer aux réglementations en vigueur et baser leurs tarifs sur des critères objectifs et justifiables. Diverses méthodes de tarification sont utilisées, telles que les modèles linéaires généralisés, les arbres de décision, les réseaux de neurones artificiels, les forêts aléatoires, les machines à vecteurs de support, les réseaux bayésiens, etc., pour analyser les données et évaluer les risques de manière précise et efficace.

Première partie

partie généralité :

0.1 Assurance :

0.1.1 Définition générale d'assurance :

L'assurance est une pratique financière qui vise à transférer le risque d'un événement incertain et potentiellement dommageable, tel qu'un accident ou une maladie, d'une personne ou d'une entreprise à une compagnie d'assurance en échange d'une prime régulière ou unique. L'objectif de l'assurance est de protéger l'assuré contre les pertes financières imprévues qui pourraient survenir en cas d'événement couvert par la police d'assurance. En cas de sinistre, l'assureur indemnisera l'assuré en fonction des modalités et des limites de la police d'assurance. Les différents types d'assurances comprennent l'assurance automobile, l'assurance habitation, l'assurance maladie, l'assurance vie, l'assurance responsabilité civile, etc.[5]

Définition juridique d'assurance :

L'article 2 de l'ordonnance n°95-07 du 25 janvier 1995 relative aux assurances définit l'assurance en référence à l'article 619 du code civil en Algérie comme suit :
- L'assurance est un contrat par lequel l'assureur s'oblige, moyennant des primes ou autres versements pécuniaires, à fournir à l'assuré ou au tiers bénéficiaire au profit duquel l'assurance est souscrite, une somme d'argent, une rente ou une autre prestation pécuniaire, en cas de réalisation du risque prévu au contrat.[8]

Définition économique d'assurance :

L'assurance est un mécanisme économique qui permet de transférer les risques de perte financière d'un individu ou d'une entreprise à une compagnie d'assurance moyennant le paiement d'une prime. Elle consiste donc à répartir le risque entre plusieurs personnes ou entités.

Plus précisément, l'assurance fonctionne selon le principe de mutualisation des risques, qui implique que les assurés paient une prime d'assurance en échange d'une garantie de la compagnie d'assurance de les indemniser en cas de sinistre. La prime est fixée en fonction du niveau de risque, c'est-à-dire de la probabilité que le sinistre se produise et du montant potentiel des pertes.
L'assurance permet ainsi à l'assuré de se prémunir contre les risques financiers

liés à un événement imprévu et coûteux (par exemple, un accident de voiture, un incendie, une maladie, etc.) en transférant ce risque à une compagnie d'assurance qui, en échange de la prime, prend en charge tout ou partie des coûts associés à ce sinistre. Cette mutualisation des risques permet aux assurés de bénéficier d'une protection financière collective plus grande que s'ils devaient supporter seuls les conséquences d'un sinistre.[1]

0.1.2 Les bases d'assurance :

L'assurance est un contrat entre une compagnie d'assurance et un assuré, dans lequel la compagnie d'assurance s'engage à payer une indemnisation en cas de réalisation d'un risque couvert par le contrat. Voici les éléments clés de l'assurance :

- Prime : l'assuré doit payer une prime à la compagnie d'assurance pour bénéficier de la couverture d'assurance. La prime est calculée en fonction du risque couvert, du niveau de couverture et des antécédents de l'assuré.

- Risque : l'assurance couvre les risques spécifiques définis dans le contrat, tels que les accidents, les maladies, les dommages matériels, etc.

- Contrat : le contrat d'assurance doit être clair sur les termes et les conditions de la couverture, y compris les limites de la couverture, les exclusions et les conditions de paiement.
- Sinistre : si le risque couvert se produit, l'assuré doit signaler le sinistre à la compagnie d'assurance et fournir les preuves nécessaires pour obtenir une indemnisation.

- Indemnisation : si le sinistre est couvert par le contrat, la compagnie d'assurance verse une indemnisation à l'assuré pour compenser les pertes subies.

En fin de compte, l'assurance permet de transférer le risque de l'assuré à la compagnie d'assurance moyennant le paiement d'une prime. Cela permet à l'assuré de se protéger contre les pertes financières en cas de réalisation d'un risque couvert par le contrat.[6]

0.1.3 Les différents types d'assurance :

Il existe plusieurs types d'assurance, qui peuvent être regroupés en quatre grandes catégories :

- L'assurance de personnes : ce type d'assurance est destiné à protéger la personne assurée et/ou sa famille contre les risques liés à la vie et à la santé. Il comprend notamment l'assurance vie, l'assurance maladie, l'assurance invalidité, l'assurance accident, etc.

- L'assurance de biens : ce type d'assurance est destiné à protéger les biens de l'assuré contre les risques liés à la propriété et à l'utilisation. Il comprend notamment l'assurance habitation, l'assurance automobile, l'assurance responsabilité civile, l'assurance des entreprises, etc.

- L'assurance voyage : ce type d'assurance est destiné à protéger l'assuré contre les risques liés aux voyages, tels que l'annulation de voyage, la perte de bagages, la maladie ou l'accident à l'étranger, etc. Il peut être souscrit pour un voyage spécifique ou pour une période plus longue.

Il est important de noter que ces catégories peuvent se chevaucher, et que certaines polices d'assurance peuvent couvrir plusieurs types de risques.[3]

0.1.4 Le rôle d'assurance :

Le rôle de l'assurance est de transférer le risque d'un événement incertain et potentiellement dommageable de l'assuré à l'assureur. En échange du paiement d'une prime, l'assureur prend en charge les risques couverts par la police d'assurance, ce qui permet à l'assuré de se protéger financièrement contre des pertes importantes en cas d'événement couvert. Le rôle de l'assurance est donc de fournir une protection financière et une tranquillité d'esprit à l'assuré en cas d'événement imprévu.

L'assurance joue également un rôle économique important en permettant la distribution des risques entre un grand nombre de personnes ou d'entreprises. En regroupant les primes de nombreux assurés, l'assureur peut couvrir les coûts des sinistres et des frais de gestion, tout en réalisant un bénéfice raisonnable. L'assurance contribue donc à la stabilité financière et à la prévention de l'insolvabilité

des entreprises et des ménages.

Enfin, l'assurance peut jouer un rôle social en aidant à réduire les coûts sociaux associés aux accidents et aux maladies, tels que les coûts de santé publique et de sécurité sociale. L'assurance peut également contribuer à la prévention des sinistres en encourageant la sécurité et la prévention des risques.[14]

0.1.5 Le contrat d'assurance :

Définition d'un contrat assurance :

Un contrat d'assurance est un accord entre un assureur et un assuré dans lequel l'assureur s'engage à couvrir certains risques en échange d'une prime payée par l'assuré.

Le contrat d'assurance peut prendre de nombreuses formes différentes en fonction du type de risque couvert et des besoins de l'assuré. Par exemple, une assurance automobile couvre les dommages causés à une voiture en cas d'accident, tandis qu'une assurance vie offre une indemnisation en cas de décès de l'assuré.

Le contrat d'assurance contient des clauses qui définissent les conditions de couverture de l'assuré. Ces conditions incluent généralement le montant de la prime, la durée de la couverture, les types de risques couverts, les exclusions et les limites de la couverture.
Il est important pour l'assuré de comprendre les termes et les conditions du contrat d'assurance avant de souscrire à une assurance. Il est également important de bien évaluer ses besoins en matière d'assurance et de comparer les offres de différents assureurs pour trouver la couverture la plus adaptée à ses besoins et à son budget.[10]

Les acteurs d'un contrat d'assurance :

Les acteurs d'un contrat d'assurance sont les personnes ou entités impliquées dans le contrat d'assurance. Voici les principaux acteurs :

- L'assuré : c'est la personne ou l'entité qui achète l'assurance pour se protéger contre un risque spécifique. Par exemple, une personne peut acheter une assurance

automobile pour se protéger contre les accidents de la route.

- L'assureur : c'est la compagnie d'assurance qui vend la police d'assurance à l'assuré et qui est responsable de couvrir les pertes en cas de sinistre. Les assureurs sont réglementés par les autorités compétentes en matière d'assurance.

- Le bénéficiaire : c'est la personne désignée par l'assuré pour recevoir les prestations de l'assurance en cas de sinistre. Par exemple, une personne peut désigner son conjoint comme bénéficiaire de son assurance vie.
- Le courtier d'assurance : c'est un professionnel qui agit en tant qu'intermédiaire entre l'assuré et l'assureur. Le courtier peut aider l'assuré à choisir la bonne police d'assurance et à négocier les termes du contrat.

- L'expert en sinistres : c'est un professionnel qui évalue les dommages causés lors d'un sinistre et qui aide l'assureur à déterminer les montants de remboursement.

- Le souscripteur : c'est la personne qui évalue le risque de l'assuré et décide d'accepter ou de refuser la demande d'assurance.[4]

0.2 Tarification :

0.2.1 Définition générale de la tarification :

La tarification est le processus de détermination des prix pour les produits ou services proposés par une entreprise. Ce processus implique l'analyse et la prise en compte de divers facteurs, tels que les coûts de production, les coûts de marketing et de distribution, la concurrence, la demande des consommateurs, la perception de la valeur du produit ou du service par les clients, ainsi que les objectifs de l'entreprise.

La tarification peut être utilisée pour influencer les comportements des clients, pour atteindre des objectifs commerciaux tels que maximiser les revenus ou la part de marché, ou pour répondre aux changements de la demande ou de la concurrence. Les entreprises peuvent utiliser différentes stratégies de tarification, telles que la tarification basée sur les coûts, la tarification dynamique, la tarification par paliers, la tarification basée sur la valeur ou encore la tarification par abonnement.

La tarification est un élément clé de la stratégie commerciale d'une entreprise et peut avoir un impact significatif sur sa rentabilité et sa compétitivité. Par conséquent, la tarification doit être soigneusement étudiée et ajustée en fonction des besoins de l'entreprise et de l'évolution du marché.[7]

Définition économique de la tarification :

En économie, la tarification désigne le processus de détermination des prix pour les biens et services proposés par une entreprise. Dans le contexte de l'assurance, la tarification consiste à déterminer la prime d'assurance que l'assuré devra payer pour obtenir une couverture d'assurance donnée.

La tarification en assurance repose sur l'évaluation du risque, c'est-à-dire la probabilité qu'un sinistre se produise et le montant potentiel des pertes. Les assureurs utilisent des modèles statistiques pour évaluer ces risques, en prenant en compte des données sur l'âge, le sexe, la profession, les antécédents médicaux, la localisation géographique, etc. de l'assuré.

Sur la base de ces données, l'assureur fixe le montant de la prime d'assurance, qui doit être suffisamment élevé pour couvrir le coût des sinistres, les frais administratifs et les bénéfices de l'assureur, tout en restant compétitif par rapport aux tarifs proposés par d'autres compagnies d'assurance.

La tarification en assurance est donc un processus complexe qui implique l'évaluation des risques, la fixation des primes, la gestion des sinistres et la concurrence entre les compagnies d'assurance.[9]

0.2.2 Les bases de la tarification :

La tarification est le processus de détermination du prix d'un produit ou d'un service offert par une entreprise. Il existe différentes méthodes de tarification, mais les bases restent les mêmes. Voici les éléments clés de la tarification :

- Coûts : la tarification doit tenir compte des coûts directs (matières premières, main-d'œuvre, etc.) et indirects (frais généraux, marketing, etc.) de la production et de la distribution du produit ou du service.

- Objectifs : la tarification doit également prendre en compte les objectifs de l'entreprise, tels que la maximisation des bénéfices, la croissance de la part de marché, la pénétration de nouveaux marchés, etc.

- Demande : la tarification doit être adaptée à la demande du marché pour le produit ou le service en question. Si la demande est élevée, les prix peuvent être plus élevés, tandis que si la demande est faible, les prix doivent être ajustés en conséquence.

- Concurrents : la tarification doit également tenir compte de la concurrence. Si les concurrents offrent des produits similaires à des prix inférieurs, l'entreprise devra peut-être ajuster ses prix pour rester compétitive.

- Segment de marché : la tarification peut varier en fonction du segment de marché ciblé. Par exemple, les prix pour les produits de luxe peuvent être plus élevés que pour les produits de base.

- Politique de tarification : l'entreprise peut choisir une politique de tarification spécifique, comme la tarification de pénétration (prix bas pour pénétrer un marché) ou la tarification de prestige (prix élevés pour les produits de luxe).

En fin de compte, la tarification doit permettre à l'entreprise de réaliser ses objectifs tout en restant compétitive sur le marché.[17]

0.2.3 Le rôle de la tarification :

Le rôle de la tarification en assurance est d'établir un prix juste et équitable pour la couverture d'un risque donné. La tarification permet aux assureurs de déterminer la prime d'assurance à facturer aux assurés pour couvrir le risque, tout en permettant à l'assureur de couvrir ses propres coûts et de réaliser un bénéfice.

Le processus de tarification implique l'évaluation du risque associé à la couverture proposée, ainsi que l'analyse des données statistiques sur les sinistres passés et les probabilités de survenance de sinistres futurs. La tarification doit également prendre en compte les coûts de gestion de l'assurance, tels que les coûts administratifs, les frais de gestion des sinistres et les coûts liés à la réassurance.

Une tarification précise est essentielle pour l'assureur pour maintenir la solvabilité financière de l'entreprise et pour offrir des primes d'assurance abordables pour les clients. Les primes doivent être suffisamment élevées pour couvrir les coûts liés aux sinistres, mais pas excessives au point de décourager les clients de souscrire à l'assurance.

La tarification est également importante pour les clients, car elle leur permet de connaître le coût de la couverture et de choisir une assurance adaptée à leurs besoins et à leur budget. Une tarification juste et transparente renforce la confiance des clients dans l'assureur et favorise la fidélité des clients à long terme.

En somme, le rôle de la tarification en assurance est d'assurer l'équilibre entre les besoins des clients et ceux de l'assureur en établissant un prix équitable pour la couverture d'un risque donné. [15]

0.3 Histoire de la tarification en assurance automobile :

L'histoire de la tarification en assurance automobile remonte au début du 20ème siècle, lorsque l'utilisation des véhicules automobiles a commencé à se généraliser. Les premières compagnies d'assurance automobile ont été créées pour couvrir les risques liés à l'utilisation de ces nouveaux moyens de transport.

Au départ, les compagnies d'assurance automobile ont fixé des primes uniformes pour tous les conducteurs, sans tenir compte de leur profil de risque individuel. Cela signifiait que les conducteurs les plus prudents et les plus expérimentés payaient les mêmes primes que les conducteurs les plus imprudents et les moins expérimentés.
Au fil du temps, les compagnies d'assurance automobile ont commencé à réaliser que cette approche uniforme ne permettait pas de refléter correctement le risque réel associé à chaque conducteur et à chaque véhicule. Elles ont donc commencé à chercher des moyens de tarification plus sophistiqués pour mieux évaluer le risque individuel et offrir des primes plus équitables.

Dans les années 1940, les compagnies d'assurance automobile ont commencé à utiliser des tables de classification de risque, qui évaluaient les risques associés

à différents types de conducteurs, tels que les jeunes conducteurs, les conducteurs expérimentés et les conducteurs avec un historique de conduite défavorable. Cela a permis aux compagnies d'assurance automobile de facturer des primes plus élevées aux conducteurs les plus à risque.

Dans les années 1970, les compagnies d'assurance automobile ont commencé à utiliser des systèmes de cotation de primes, qui évaluaient le risque individuel en fonction de plusieurs facteurs, tels que l'âge, le sexe, l'historique de conduite, le type de véhicule et la localisation géographique. Cela a permis aux compagnies d'assurance automobile de mieux évaluer le risque individuel et de proposer des primes plus personnalisées.

Aujourd'hui, les compagnies d'assurance automobile continuent à utiliser des systèmes de cotation de primes sophistiqués, qui prennent en compte une grande variété de facteurs pour évaluer le risque individuel et offrir des primes personnalisées. Les avancées technologiques, telles que l'utilisation de l'analyse des données et de l'apprentissage automatique, ont permis aux compagnies d'assurance automobile de développer des modèles de tarification encore plus précis et sophistiqués.[2]

0.4 La tarification en assurance automobile :

0.4.1 La définition de la tarification en assurance automobile :

La tarification en assurance automobile est le processus par lequel une compagnie d'assurance calcule le coût de la prime d'assurance automobile pour un client. Elle prend en compte plusieurs facteurs, tels que l'âge, le sexe, l'expérience de conduite, le type de véhicule, la fréquence d'utilisation, le lieu de résidence et l'historique des sinistres.

La tarification en assurance automobile peut également inclure des facteurs tels que la couverture d'assurance demandée, le niveau de franchise et les rabais applicables. Les rabais peuvent être offerts pour des choses comme la possession d'un système antivol ou la participation à des cours de conduite défensive.

La tarification en assurance automobile vise à établir une prime qui reflète le risque que représente un client pour la compagnie d'assurance. Les clients qui

sont considérés comme présentant un risque plus élevé, comme les conducteurs novices ou les conducteurs avec un historique de sinistres, paieront généralement des primes plus élevées que les conducteurs expérimentés qui ont un bon dossier de conduite.[13]

0.4.2 La relation entre automobile et tarification assurance :

La relation entre automobile et tarification assurance est étroite car les compagnies d'assurance automobile utilisent de nombreux facteurs pour évaluer le risque et déterminer le montant des primes d'assurance à facturer aux conducteurs.

Certains des facteurs clés que les compagnies d'assurance automobile prennent en compte dans la tarification de l'assurance comprennent :

Le profil du conducteur - cela comprend l'âge, le sexe, l'état civil, l'expérience de conduite et l'historique des accidents. Les conducteurs plus jeunes ou ceux qui ont eu des accidents de voiture antérieurs peuvent être considérés comme étant plus à risque, ce qui peut entraîner des primes plus élevées.

Le type et le modèle de la voiture - les voitures de sport ou les voitures de luxe peuvent coûter plus cher à assurer car elles sont souvent plus chères à réparer ou à remplacer en cas de dommages.

L'utilisation de la voiture - les voitures qui sont utilisées pour les trajets quotidiens ou les longues distances peuvent être considérées comme plus à risque, tandis que les voitures qui sont utilisées occasionnellement ou pour des loisirs peuvent avoir des primes d'assurance plus basses.

La zone géographique - les tarifs d'assurance peuvent varier selon l'endroit où vous vivez et conduisez. Les zones avec des taux d'accidents plus élevés peuvent entraîner des primes d'assurance plus élevées.

Les antécédents de sinistre - si un conducteur a un historique de sinistres ou de réclamations d'assurance, cela peut augmenter le risque perçu par la compagnie d'assurance et entraîner des primes plus élevées.

En fin de compte, plus le risque perçu par la compagnie d'assurance est élevé, plus la prime d'assurance sera élevée. Les conducteurs peuvent souvent économiser de l'argent sur leur assurance automobile en magasinant autour pour trouver les meilleures offres et en choisissant des options de couverture plus appropriées pour leurs besoins individuels.[16]

0.4.3 Les types de la tarification en assurance automobile :

Il existe plusieurs types de tarification en assurance automobile. Voici les principaux :

- La tarification individuelle : elle est basée sur les caractéristiques spécifiques de chaque conducteur, telles que l'âge, l'expérience de conduite, l'historique des sinistres, le type de véhicule, la fréquence d'utilisation et le lieu de résidence. Elle permet de personnaliser la prime d'assurance en fonction des risques individuels.

- La tarification de groupe : elle est appliquée à des groupes de conducteurs partageant des caractéristiques similaires, telles que les jeunes conducteurs, les conducteurs âgés, les conducteurs avec un dossier de conduite impeccable, etc. Elle permet de proposer des primes d'assurance plus avantageuses pour ces groupes.

- La tarification territoriale : elle prend en compte le lieu de résidence du conducteur pour évaluer le risque d'accident. En général, les zones urbaines sont considérées comme présentant un risque plus élevé que les zones rurales, et les primes d'assurance peuvent donc être plus élevées en conséquence.
- La tarification en fonction de l'utilisation : elle est basée sur la fréquence d'utilisation du véhicule. Les conducteurs qui utilisent leur voiture régulièrement ont généralement des primes d'assurance plus élevées que ceux qui ne l'utilisent que de manière occasionnelle.

- La tarification mixte : elle combine plusieurs de ces types de tarification pour proposer des primes d'assurance plus précises et adaptées aux besoins individuels de chaque conducteur.[19]

0.4.4 Les éléments de la tarification en assurance automobile :

Les éléments de tarification en assurance automobile peuvent varier selon les compagnies d'assurance et les réglementations en vigueur dans chaque pays, mais

voici quelques-uns des éléments les plus couramment utilisés :

*'code agence' : Le code unique attribué dans compagnie d'assurance.
*'Nom agence' : Le nom de l'agence d'assurance automobile.
*'code DR' : Le code unique attribué à une direction régionale.
*'Nom DR' : Le nom de la direction régionale.
*'produit' : Le produit d'assurance automobile associé à la police d'assurance.
*'garantie' : La garantie spécifique ou le type de couverture fourni par la police d'assurance.
*'branche' : La branche d'assurance à laquelle la police d'assurance appartient (par exemple, assurance automobile, assurance risque simple et divers, etc.).
*'Nume police' : Le numéro de police unique attribué à la police d'assurance.
*'Nume avenant' : avenant est utilisé pour désigner un document qui modifie ou complète un contrat d'assurance.
*'ID' :Un identifiant unique pour chaque assuré,contient un numero de police,code d'agence et num d'avenant.
*'Annee' : L'année de souscription ou de validité de la police d'assurance.
*'mois' : Le mois de souscription ou de validité de la police d'assurance.
*'Prime' : Le montant de la prime d'assurance.
*'DATENAISS' : La date de naissance de l'assuré principal ou du conducteur.
*'SEXE CONDUCTEUR' : Le sexe du conducteur principal ou de l'assuré principal.
*'MARK VOITURE' : La marque du véhicule assuré.
*'TYPE VOITURE' : Le type ou le modèle du véhicule assuré.
*'CODE USAGE' : Le code qui représente l'utilisation prévue du véhicule assuré (par exemple, usage personnel, usage professionnel, etc.).
*'CODE ZONE' : Le code de la zone géographique dans laquelle le véhicule est principalement utilisé.
*'PUISS VOITURE' : La puissance du moteur du véhicule assuré.
*'DATE miseCIRCULA' : La date de la première mise en circulation du véhicule.
*'NUMEIMMA' : Le numéro d'immatriculation du véhicule.
*'DATE DELIV PERMIS' : La date de délivrance du permis de conduire du conducteur principal.
* 'TYPE PERMIS' : Le type de permis de conduire du conducteur principal.
* 'COD NAT VE' : Le code qui représente la nature du véhicule (par exemple, véhicule particulier, véhicule utilitaire, etc.).

Deuxième partie

Partie théorique :

Chapitre 1

Chapitre 1 : Les modèles linéaires généralisés :

En statistique, la régression linéaire consiste à décrire la variable Y avec les variables explicatives $X_1,...,X_p$ comme suit :[12]

$Y = X\beta + \varepsilon$

Dans cette formule :

- Y représente la variable dépendante que nous cherchons à prédire.
- X est une matrice qui représente les variables explicatives.
- β est un vecteur de coefficients à estimer pour chaque variable explicative.
- ε est un terme d'erreur ou résidu qui représente l'écart entre la prédiction ($X\beta$) et la valeur réelle de Y.

avec les hypothèses fortes suivantes :

— Les variables Y et ε sont indépendantes .

— $Y|X \sim \mathcal{N}(X\beta, \sigma)$ et $\varepsilon|X \sim \mathcal{N}(0, \sigma)$

La notation $Y|X \sim \mathcal{N}(X\beta, \sigma)$ signifie que la variable Y, conditionnée par les valeurs de X, suit une distribution normale (gaussienne) avec une moyenne de $X\beta$ et une variance de σ. Cela signifie que pour chaque combinaison de valeurs de X, la variable Y suit une distribution normale dont la moyenne est déterminée par $X\beta$ (produit matriciel de X et β) et la variance est σ.

De même, la notation $\varepsilon|X \sim \mathcal{N}(0, \sigma)$ signifie que le terme d'erreur ε, condi-

tionné par les valeurs de X, suit une distribution normale avec une moyenne de 0 et une variance de σ. Cela indique que pour chaque combinaison de valeurs de X, le terme d'erreur suit une distribution normale centrée autour de zéro avec une variance constante de σ.

On cherche alors à trouver β tel que E[Y| X] = Xβ (cette formulereprésente l'espérance conditionnelle de la variable dépendante Y étant donné les valeurs des variables explicatives X, dans le cadre de la régression linéaire tel que :E[Y|X] est l'espérance conditionnelle de la variable Y, ce qui signifie qu'il s'agit de la valeur moyenne attendue de Y pour chaque combinaison de valeurs de X. Cela représente la meilleure prédiction possible de la valeur de Y connaissant les valeurs de X.

Xβ est le produit matriciel des variables explicatives X et des coefficients β. Les coefficients β sont les paramètres estimés dans la régression linéaire qui déterminent l'impact de chaque variable explicative sur la variable dépendante Y. Le produit Xβ représente la contribution combinée des variables explicatives à la prédiction de Y) soit la meilleure prédiction possible de Y connaissant X, au sens des moindres carrés. La variable ε, appelée résidu, correspond alors a l'écart entre la prédiction et la valeur réelle.

Comme mentionné précédemment, les modèles linéaires classiques ont l'avantage d'être faciles à mettre en œuvre, mais ils supposent que la variable qu'ils décrivent suit une distribution gaussienne, ce qui n'est pas le cas en assurance. Ainsi, la méthode du modèle linéaire généralisé est une extension du modèle linéaire classique, permettant de modéliser des phénomènes dont les distributions ne sont pas gaussiennes mais étendues à une famille spécifique de lois appelée famille exponentielle. Ainsi, ces modèles permettent de mieux appréhender les réalités de l'assurance.

1.0.1 Définition de GLM :

GLM signifie "Generalized Linear Model" en anglais, que l'on peut traduire en français par "modèle linéaire généralisé". Il s'agit d'une famille de modèles statistiques qui permet de modéliser une variable de réponse en fonction d'une ou plusieurs variables explicatives, en utilisant une fonction de lien non linéaire pour la relation entre la moyenne de la variable de réponse et les prédicteurs.

Le GLM est une généralisation du modèle linéaire classique qui permet de

modéliser des variables de réponse qui ne suivent pas une distribution normale, par exemple des variables binaires, des compteurs (Poisson), des variables continues positives (gamma), ou des variables continues qui ne sont pas normalement distribuées. Le GLM permet également d'inclure des variables explicatives continues ou catégorielles, et des interactions entre les variables explicatives.

Le modèle linéaire généralisé est largement utilisé en sciences sociales, en sciences de la vie, en écologie, en économie, en finance, en marketing, et dans de nombreux autres domaines où l'on souhaite modéliser une variable de réponse en fonction de plusieurs prédicteurs.

1.0.2 Principe et hypothéses des GLM :

Le modéle linéaire généralisé consiste à exprimer, par un prédicteur linéaire, l'influence des variables $X_1, \ldots, X_p$ sur une fonction (appelée fonction de lien) de l'espérance de Y (la variable à expliquer).

Autrement dit, en ajustant les paramètres, nous pouvons explorer la relation entre la variable de réponse et un ensemble de variables explicatives β_i Son équation est la suivante :

$\eta = g(E[Y|X]) = X\beta$

On cherche donc à modéliser η sous les hypothéses d'écrites ci-dessous.
Trois hypotéses permettent de caractériser un modéle linéaire généralisé :

- **Hypothése 1 : La distribution de la variable à expliquer**

La variable réponse Y est une variable aléatoire représentant le phénomène que nous essayons d'expliquer, et nous connaissons sa valeur attendue. Les observations yi de ces variables sont dites des réalisations indépendantes obéissant à des lois de probabilité appartenant à la famille exponentielle.

- **Définition : Famille exponentielle**

Soit X une variable aléatoire. X appartient à la famille exponentielle si sa densité par rapport à la mesure dominante peut s'écrire sous la forme :

f_X(x, θ) = a(x)b(θ) exp[$\wedge$(θ).T(x)]
ou a(.),b(.), $\wedge$ (.) et T(.) sont des fonctions mesurables.
Dans cette définition, la mesure dominante est la combinaison de la mesure de Lebesgue pour la loi continue et de la mesure de Dirac pour la loi discrète. La densité de cette famille peut aussi s'écrire sous la forme :

$$\underset{\theta,\phi(x)}{f} = \exp\left(x\theta - \frac{b(\theta)}{a(\phi)} + c(x,\phi)\right)$$

où $a(\cdot)$ et $c(\cdot)$ sont des fonctions dérivables, et $b(\cdot)$ est de classe C^3, de dérivée première inversible. θ est appelée paramètre d'intérêt et Φ paramètre de dispersion.

Donc, sous ce script se trouvent les propriétés de base de la famille exponentielle :

$$\left\{\begin{matrix} E[X] = b'(\theta) \\ V[X] = b''(\theta)a(\phi) \end{matrix}\right\}$$

La famille exponentielle englobe la plupart des lois standards, ce qui offre un trés grand champ d'application.
Elle comprend en effet :
• La loi normale
• La loi de poisson
• La loi gamma
• La loi exponentielle
• La loi binomiale
• La loi géométrique
• La loi binomiale négative

- **Hypothése 2 : L'expression de la linéarité**

η(X) = Xβ est défini comme un prédicteur linéaire ou β est un vecteur de paramétres inconnus de taille p et X une matrice N × p connue, fixée par l'expérience.

Les coefficients $\underset{i}{\beta}$ doivent etre estimés.

- **Hypothése 3 : Le lien entre la variable à expliquer et les variables explicatives**

Le principe au cœur du GLM consiste à appliquer une transformation de maniére à ce que η et X soient liées par une relation linéaire.
Par conséquent, la valeur attendue de Y connaissant X doit dépendre du prédicteur linéaire via la fonction de couplage g(). Il est réel, monotone et différentiable.
Cette fonction de lien doit vérifier que l'ensemble de définition de la fonction, c'est-à-dire $g^{-1}(R)$, coincide avec les valeurs possibles de la variable réponse.

Chacune des lois de probabilités de la famille exponentielle posséde une fonction de lien spécifique, dite "fonction de lien canonique". Il s'agit de la fonction qui lie la moyenne de Y au paramétre d'intérét Φ On cherche donc g telle que :

g(E(Y)) = Φ g(.) = (b')$^{-1}$(.)

Le tableau suivant répertorie les fonctions de lien canonique les plus courantes :

Loi de probabilité	Fonction de lien canonique	
Bernouilli	η =ln(μ/(1-μ))	Logit
Binomiale négative	η = ln(p)	Log
Gamma	η =(1/ μ)	Inverse
Poisson	η = ln(μ)	Log
Normale	η= θ	Identité

TABLE 1.1 – Tableau des fonctions de lien usuelles

Bien que la fonction de lien standard soit statistiquement la plus adaptée à la distribution des réponses, vous pouvez choisir une autre fonction de lien qui satisfait aux conditions ci-dessus pour répondre à vos critères spécifiques, en particulier la structure de votre modèle.

Cela s'écrit :

$E(Y \mid X) = g^{-1}(X\beta)$

L'effet de chaque coefficient β dépend donc de la fonction choisie. En pratique, les fonctions identité et logarithmique sont privilégiées. En effet, le premier propose un modèle additif alors que le second propose un modèle multiplicatif. Ces deux fonctions facilitent donc l'arbitrage et le réglage des coefficients.

1.0.3 Estimation des coefficients du modéle :

Aprés avoir choisi la distribution de la variable à expliquer Y et la fonction de lien déterminant la structure du modéle, il reste à estimer les coefficientsβ_i ainsi que le paramétre de dispersion, qui sont inconnus.
Cette étude utilise la méthode du maximum de vraisemblance, qui est implémentée par défaut dans la plupart des logiciels.

Revenons d'abord au sens du mot "probabilité". La probabilité mesure l'adéquation entre une distribution observée dans un échantillon aléatoire et une loi de probabilité destinée à imprimer la réalité sur la population à partir de laquelle l'échantillon a été tiré. Nous connaissons la fonction de densité de la loi de probabilité théorique qui correspond le mieux à la population, mais nous ne connaissons pas ses paramètres, nous devons donc utiliser des statistiques d'échantillons pour l'estimer. La probabilité est donc l'une des techniques permettant de trouver les estimateurs les plus pertinents. Nous supposons que la probabilité d'observer un échantillon est le produit des probabilités d'observer chaque réalisation.

Ainsi, pour n observations indépendantes $(y_1, \ldots, y_n)$ de la variable réponse Y, de loi appartenant à la famille exponentielle, de paramètres θ_i et de densité f, la vraisemblance s'écrit :

$L(Y, \theta, \phi) = \prod_{i=1}^{n} f_{\theta,\phi}(y_i)$

En fait, il est plus pratique de prendre le logarithme de cette fonction pour maximiser la probabilité. Cela transforme le produit des densités en une somme, facilitant les calculs suivants : Puisque la fonction logarithmique est strictement croissante, le maximum de la fonction de vraisemblance est aussi le maximum de la log-vraisemblance.

En remarquant que θ dépend de β, on peut écrire la log-vraisemblance comme

suit :

$$\mathrm{l}(\beta) = \sum_{i=1}^{n} \ln(f_{\theta,\phi}(y_i)) = \sum_{i=1}^{n} \left(y_i\theta - \frac{b(\theta)}{a(\phi)+c(y_i,\phi)} \right) = \sum_{i=1}^{n} y_i$$

La formule donnée est une représentation abrégée de la fonction de vraisemblance (likelihood) dans le contexte des modèles linéaires généralisés. La fonction de vraisemblance mesure la probabilité d'observer les données réelles (y) en fonction des paramètres du modèle (β).
La formule est composée de trois termes dans la somme, chacun correspondant à une étape spécifique dans le calcul de la fonction de vraisemblance.

Le premier terme,$\ln(f_{\theta,\phi}(y_i))$, représente le logarithme de la fonction de densité de probabilité (f) évaluée pour la variable $y_i, paramétriséepar\theta$ et ϕ. Cette fonction de densité de probabilité est spécifiée dans le cadre des modèles linéaires généralisés.
Le deuxième terme,$\left(y_i\theta - \frac{b(\theta)}{a(\phi)+c(y_i,\phi)} \right)$, représente une transformation spécifique appliquée à $y_i, et\phi$. Les fonctions b, a et c sont des fonctions spécifiques au modèle linéaire généralisé utilisé.

Le troisième terme, $(y_i), représentesimplementlavariable(y_i)elle-même.$

Pour obtenir les équations de vraisemblance, on Calcule :

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

Comme :

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\sigma)} = \frac{y_i - \mu_i}{a(\sigma)}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{Var(Y_i)}{a(\sigma)} = b''(\theta_i)$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}, \quad \text{car } \eta_i = \sum_j \beta_j x_{ij}$$

Puisque, $\frac{\partial \mu_i}{\partial \eta_i}$ dépends de la fonction lien $\eta_i = g(\mu_i)$ du modèle alors :

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \times \frac{a(\phi)}{Var(Y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij}$$

Les derivés de chaque terme sont :

$\frac{\partial l_i}{\partial \theta} = y_i - \frac{b'(\theta)}{\alpha(\phi)} = y_i - \frac{\mu_i}{\alpha(\theta)}$

$\frac{\partial \theta}{\partial \mu_i} = \frac{\partial^2 b(\theta)}{\partial \theta^2} = \frac{V[Y_i|X]}{\alpha(\theta)}$

$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = (g^{-1})'(\eta_i)$

$\frac{\partial \eta}{\partial \beta_j} = \frac{\partial X_i \beta}{\partial \beta_j} = x_{ij}$

Dans ce cas, l'équation de probabilité, appelée équation de Wedderburn, s'écrit :

$\sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{v[Y_i|X]} \times x_{ij} \times \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{pour } j = 1, \ldots, p$

Malheureusement, ces équations ne sont pas linéaires en β et nous ne connaissons pas $\underset{i}{\mu}$.
Pour trouver la solution, nous devons utiliser un algorithme itératif avec une matrice Hessienne.
Cette étude utilise l'algorithme de Newton-Raphson. Ce processus est :

Etape 1 :Initialisation

On part d'une valeur initiale $\underset{0}{\beta}$ de β.

Etape 2 :Itérations

On itére selon la procédure de récurrence suivante :

$\underset{k+1}{\beta} = \underset{k}{\beta}\text{-H}\nabla$

où $\text{H} = \left(\frac{\partial^2 l}{\partial \beta_i \partial \beta_j}\right)_{i,j=1,\ldots,p}$ est la matrice Hessienne, et $\nabla = \left(\frac{\partial l}{\partial \beta_j}\right)_{j=1,\ldots,p}$ le vecteur score.

Etape 3 : Arrêt

On arrête le processus lorsqu'on observe une convergence vers l'ensemble des β optimaux, i.e. lorsqu'une nouvelle itération de l'algorithme n'améliore pas plus la vraisemblance (par défaut, amélioration $\prec 0.01\%$).

En posant W matrice diagonale de terme général $w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / V(y_i)$, l'algo-

rithme utilise comme approximation :

E[H] = -$\phi^t XWX$

LIMITATIONS :

Le modèle GLM est dit paramétrique, en effet il nécessite de préciser une loi pour la variable d'intérêt Y| X=x. Ici, nous avons choisi la distribution de Poisson. Le modèle est plus linéaire, de même que l'effet des variables explicatives. Pour atténuer cette hypothèse, nous pouvons nous tourner vers les modèles additifs généralisés (GAM).
GLM ne peut pas modéliser des effets différents pour la même variable explicative. Les mêmes coefficients s'appliquent aux variables continues. Il a une forme monotone. Aussi, la gestion des valeurs aberrantes et des valeurs manquantes est délicate.
Enfin, la modélisation des interactions entre variables est possible, mais comme la sélection en amont, relève souvent du jugement d'un expert. En particulier, il peut être intéressant que la variable exclue dans la méthode basée sur l'AIC soit liée à une autre variable. Les modèles peuvent prendre beaucoup de temps à s'exécuter et nous n'avons pas le luxe de tester toutes les interactions possibles et de maintenir un modèle optimal.
Pour toutes ces raisons, nous nous tournons maintenant vers une nouvelle méthode non paramétrique : les arbres de régression.[18]

1.0.4 Le XGBoost :

Définition de XGBoost :

XGBoost (eXtreme Gradient Boosting) est un algorithme d'apprentissage automatique de type "boosting" pour la régression et la classification. Il a été développé par Tianqi Chen en 2014 et est devenu très populaire en raison de sa haute performance et de son efficacité dans un large éventail de problèmes de prédiction.

Le principe de XGBoost est de construire un modèle prédictif en combinant plusieurs modèles plus simples, appelés arbres de décision. À chaque étape, l'algorithme essaie d'ajouter un nouvel arbre de décision qui améliore la précision globale du modèle. La particularité de XGBoost est qu'il utilise une fonction de

coût spéciale pour évaluer la performance de chaque nouvel arbre et choisir le meilleur à ajouter au modèle.

XGBoost utilise également des techniques d'optimisation pour améliorer la vitesse et la précision de l'apprentissage, telles que le calcul parallèle, la réduction de la dimensionnalité et la régularisation pour éviter le surapprentissage. Il est souvent utilisé pour des problèmes de classification et de régression dans des domaines tels que la finance, l'industrie et la recherche en sciences sociales.[11]

Le Boosting :

Une seconde façon d'améliorer les arbres est le boosting. Tout comme le bagging, le boosting repose sur un système additif d'arbres. A l'étape t le modèle est

$$\hat{yi}^{(t)} = \sum_{k=1}^{t} \nu f_k(xi) = \hat{yi}^{(t-1)} + \nu f_t(xi)$$

Mais les arbres ne sont pas indépendants, ils sont construits de façon récursive, l'étape t apprend des erreurs de l'étape t-1.
ν est un hyper-paramètre qui permet de limiter volontairement l'apprentissage de l'étape t pour ne pas faire de sur-apprentissage et laisser une chance aux futurs itérations d'apprendre. Dans la suite se paramètre sera fixé à 1 pour faciliter la lecture.
La fonction objectif Obj à pour but de juger de la qualité du modèle, elle est souvent composée de deux termes, l'un juge la qualité des prédictions l et l'autre pénalise la complexité du modèle Ω.
Obj(Θ) = l(θ) + $\Omega(\Theta)$

Dans le cas du boosting d'arbres à l'étape t,

$$\text{Obj}^{(t)} = \sum_{k=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

nétant le nombre d'observation à disposition.
Il est possible de réécrire cette équation en séparant l'arbre de l'étape t de ceux des étapes précédentes.

$$\text{Obj}^{(t)} = \sum_{k=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(xi)) + \Omega(f_t) + \sum_{k=1}^{t-1} \Omega(f_k)$$

La problématique est donc de choisir $\underset{t}{f}$ de façon à minimiser cette fonction objectif

Le gradient boosting :

Pour résoudre ce problème d'optimisation, la méthode de gradient boosting approxime l par décomposition de Taylor.

Le XGBoost :

XGBoost est un algorithme connu pour ses performances et son temps de prédiction et de calcul.
XGboost utilise une approximation de second ordre de la fonction objectif.

$$\begin{cases} \underset{i}{g} = \underset{\hat{y}i^{(t-1)}}{\partial} \underset{i}{l}(y, \underset{i}{\hat{y}}^{(t-1)}) \\ \\ \underset{i}{h} = \underset{\hat{y}i^{(t-1)}{}_2}{\partial} \underset{i}{l}(y, \underset{i}{\hat{y}}^{(t-1)}) \end{cases}$$

Pour rappel, l'approximation d'ordre 2 en série de Taylor d'une fonction f est $f(x + \Delta x) = f(x) + f'(x)\Delta x + 1/2 f''(x)\Delta x^2$

L'approximation d'ordre 2 de la fonction objectif est donc,

$$\text{Obj}^{(t)} \approx \sum_{k=1}^{n} [\underset{i}{l}(y, \underset{i}{\hat{y}}^{(t-1)}) + \underset{i}{g}.\underset{t}{f}(\underset{i}{x}) + 1/2 \underset{i}{h}.\underset{t}{f}(\underset{i}{x})^2] + \Omega(\underset{t}{f}) + \sum_{k=1}^{t-1} \Omega(\underset{k}{f})$$

Soit les notations suivantes :

- $\mathbf{w}_q(x) = \mathbf{f}_t(x)$, où q représente la structure de l'arbre et $\mathbf{w}_q(x)$ est le poids donné par l'arbre q à x.

-$\underset{j}{I} = \{ \text{i} | \text{q}(\underset{i}{x}) = j \}$ l'ensemble des individus dans la feuille j de l'arbre q Dans l'algorithme de l'XGBoost

$\Omega(\underset{t}{f}) = \gamma \mathrm{T} + 1/2\lambda \sum_{j=1}^{T} \|\underset{j}{w}\|^2$

Pour un arbre contenant T feuilles.
En omettant les termes non liés à l'étape t est obtenu,

$\mathrm{Obj}^{(t)} \approx \sum_{k=1}^{n} [\underset{i}{g}.\underset{t}{f}(\underset{i}{x}) + 1/2\underset{i}{h}.\underset{t}{f}(\underset{i}{x})^2] + \gamma \mathrm{T} + 1/2\lambda \sum_{j=1}^{T} \|\underset{j}{w}\|^2$

La première somme peut être réécrite en indexant les feuilles. En effet, à la fin de l'arbre, tous les individus se trouvent dans l'une des feuilles. Supposons que nous ayons T dans l'arbre.aller

$\mathrm{Obj}^{(t)} \approx \sum_{j=1}^{T} [(\sum_{i \in \underset{j}{I}} \underset{i}{g}).\underset{j}{w} + 1/2(\sum_{i \in \underset{j}{I}} \underset{i}{h} + \lambda).\overset{2}{\underset{j}{w}}] + \gamma \mathrm{T}$

Les deux résultats suivants sont des résultats classiques pour des fonctions polynomiales de degré 2.

$$\begin{cases} \mathrm{argmin}_x \mathrm{Gx} + 1/2\mathrm{Hx}^2 & = \\ \qquad -G/H, H \succ 0 & \\ \min\limits_x Gx + 1/2Hx^2 & = \\ \qquad -G^2/H & \end{cases}$$

En notant

$$\begin{cases} \underset{i}{G} = \sum_i \in \underset{j}{I} \underset{i}{g} \\ \\ \underset{i}{h} = \sum_i \in \underset{j}{I} \underset{i}{h} \\ \\ \end{cases}$$

$\mathrm{Obj}^{(t)} \approx \sum_{j=1}^{T} [\underset{i}{G}.\underset{j}{w} + 1/2(\underset{i}{H} + \lambda).\underset{j}{w^2}] + \gamma \mathrm{T}$

En fixant la structure de l'arbre q alors les poids optimaux sont

$\mathrm{Wj}^* = -\underset{j}{G}/\underset{j}{H} + 1\, j \in \{\ 1,.....,\mathrm{T}\ \}$

Et la fonction objective optimale est à une constante près,

obj= -1/2$\sum_{j=1}^{T} \underset{j}{G^2}/(\underset{j}{H} + \lambda) + \gamma \mathrm{T}$

Il faut donc choisir la structure de l'arbre q pour décider. Pour ce faire, la

même méthode que pour les arbres est utilisée. Un arbre avec un nœud racine simple est calibré, puis pour chaque feuille le nœud suivant est calibré et le gain est optimisé par rapport à la fonction objectif Obj. Cette victoire est

$$\text{Gain} = \underset{g}{G}^2/(\underset{g}{H}+\lambda) + \underset{d}{G}^2/(\underset{d}{H}+\lambda) - (\underset{g}{G}+\underset{d}{G})^2/(\underset{g}{H}+\underset{d}{H}+\lambda) - \gamma$$

S'il n'y a pas de coupures permettant des gains strictement positifs, l'algorithme s'arrête.

L'algorithme XGBoost utilise la mémoire et la puissance de traitement de manière très efficace. Pour plus d'informations sur les techniques d'optimisation numérique et matérielle, voir l'article " XGBoost : A Scalable Tree Boosting System " [Chen et Guestrin, 2016]. [11] **Résumé des paramètres usuels**

— $\nu \in [0,1]$ qui limite l'apprentissage des différentes étapes pour laisser les suivantes apprendre
— la profondeur maximum des arbres
— le nombre de pas
— $\gamma \in [0, \infty]$ qui pénalise le gain et donc favorise un arrêt prématuré de l'algorithme
— le nombre de variables explicatives sélectionnées par itération
— le nombre d'observations sélectionnées par itération
Ces paramètres peuvent être calibrés par validation croisée

Chapitre 2

Chapitre 2 :Classification des résidus des modéles :

Définition sur maching learning (apprentissage automatique) : L'apprentissage automatique, également connu sous le nom de "machine learning" en anglais, est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre et de s'améliorer à partir de données, sans avoir besoin d'être explicitement programmés. Cela se fait en utilisant des algorithmes et des modèles mathématiques pour identifier des modèles dans les données et créer des prédictions et des décisions basées sur ces modèles. L'apprentissage automatique est utilisé dans de nombreux domaines, tels que la reconnaissance de la parole, la vision par ordinateur, la prédiction de fraudes, la recommandation de produits et la personnalisation de l'expérience utilisateur.

La deuxième étape consiste à classer les compartiments des véhicules dans différentes classes de risque en fonction des résidus détectés. Par conséquent, nous voulons distinguer les types de véhicules " plus risqués " associés à un carnet de commandes moyen élevé des types de véhicules " plus fiables" en fonction uniquement de leurs caractéristiques. H. Sans tenir compte d'autres facteurs tels que le comportement du conducteur. C'est pourquoi nous nous présentons dans le cadre d'un apprentissage supervisé.

Nous avons choisi d'utiliser l'une des méthodes d'apprentissage automatique les plus intuitives et les plus populaires, l'algorithme CART. ça a l'avantage Des résultats faciles à interpréter peuvent être produits grâce à la représentation graphique de l'arbre et des règles de décision explicites. [12]

2.0.1 L'algorithme CART :

Principe des CART

Le but de la méthode CART est de segmenter la population selon un ensemble de variables explicatives en divisant à plusieurs reprises les données en deux groupes. Son algorithme est basé sur la construction d'arbres de décision avec des partitions récursives binaires. Il s'agit d'une séquence de nœuds sur 0, 1 ou 2 branches. Une feuille est un nœud final, c'est-à-dire un nœud où une branche ne commence pas. Un nœud a la particularité d'être défini par une combinaison de deux éléments principaux.

— Une variable parmi toutes les variables explicatives.
— Une subdivision résultant en une subdivision en deux classes. Il consiste à déterminer le seuil ou la classification des modalités pour des variables quantitatives sélectionnées Deux phrases si la variable est qualitative.
Chaque nœud contient donc un sous-ensemble de la population qui est supposé devenir de plus en plus homogène au fur et à mesure de l'itération (c'est-à-dire vers le bas de l'arbre).
A titre d'exemple, prenons une variable de réponse quantitative Y et deux variables explicatives discrètes X1 et X2 et considérons l'arbre suivant :

Chaque nœud peut se voir attribuer une question binaire oui ou non.
Obtenez le premier nœud (appelé la racine). Divisez la première population en deux parties.
Sous-population par variable explicative X1 considérée comme la plus discriminante. Pour affecter chaque individu à sa sous-population, on pose la question La variable X1 est-elle égale à 1 ? Si la réponse est oui, la personne suit la branche de gauche menant au deuxième nœud. En revanche, si la réponse est non, il sera affecté à la feuille 3 car il prend la branche de droite.

La raison est la même pour le deuxième nœud.
L'avantage d'une telle représentation est la lisibilité apportée par la stratification de l'effet de la variable dérivée sur la variable réponse. Donc la première question est la plus importante, la seconde est la moins importante, et ainsi de suite.
Dans cet exemple, nous nous retrouvons avec un arbre à trois feuilles. Cela signifie que l'algorithme CART divise les observations en trois groupes homogènes. Les règles qui les définissent sont prises en fonction des variables explicatives. En partant de la racine, suivez les chemins découverts à travers les questions posées

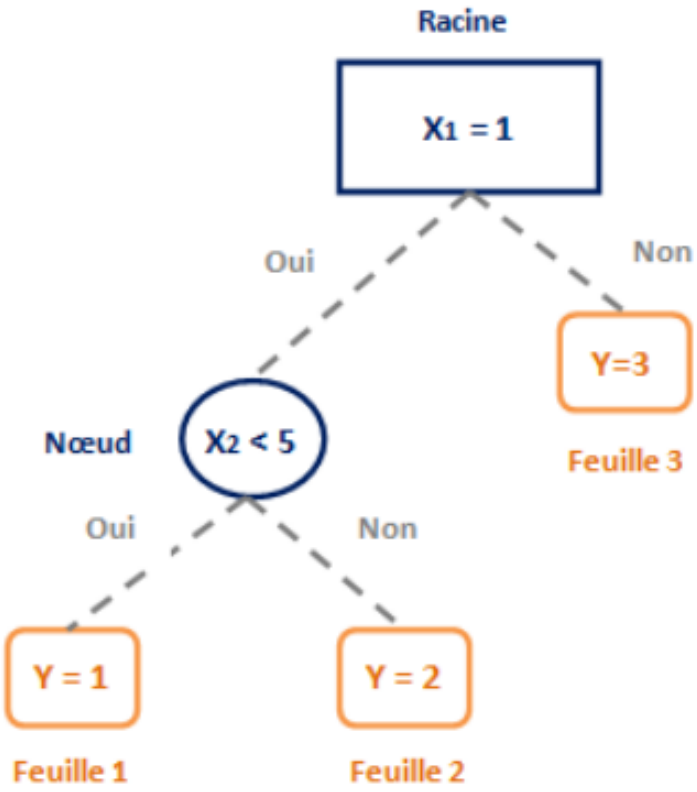


FIGURE 2.1 – Exemple d'arbre d'écision

au niveau du nœud, et suivez ceci jusqu'aux feuilles correspondantes.
Les règles de classement pour chaque joueur sont les suivantes :
— Si X1 = 1 et X2 $\prec$ 5 alors l'individu appartient au groupe 1.
— Si X1 = 1 et X2 $\succeq$ 5 alors l'individu appartient au groupe 2.
— Si X1 $\neq$ 1 alors l'individu appartient au groupe 3.
La valeur prédite de Y affectée au groupe est lue à chaque fois au niveau feuille correspondant. Il s'agit de la moyenne des Y observations pour l'individu. appartiennent à cette feuille.

c $\hat{f}$ de E[Y|X] = f(X) tel que, pour une observation x :

$\hat{f}(x) = \sum_{j=1}^{3} \times \underset{x \in F j}{1}$

où $\underset{x \in Fj}{1}$ est la fonction indicatrice associée à la feuille Fj et $\bar{Y}j$ désigne la moyenne empirique dans le groupe j.

La subtilité du CART réside dans sa recherche de modèles économiques. Un arbre trop détaillé, associé à une surparamétrisation, est instable et plus susceptible de ne pas prédire de nouvelles observations. A l'inverse, un arbre avec trop peu de branches produira un modèle imprécis. Par conséquent, nous essayons de faire un arbre robuste.
Sa taille représente un compromis entre les deux enjeux évoqués.
Trouver cet arbre optimal implique d'appliquer deux phases clés :
- Phase 1 :Construction de l'arbre maximal
Distribue les informations au sein d'un nœud selon des critères de distribution spécifiques.
— Phase 2 : Elagage de l'arbre
Sélectionnez le meilleur sous-arbre parmi les critères d'arrêt.

Phase 1 : Construction de l'arbre maximal

La clé de cette phase consiste à décider quelles questions placer à chaque niveau de nœud. Vous devez sélectionner les variables et les départements qui les définissent. La croissance des arbres se déroule comme suit :
1. L'algorithme commence par choisir les variables explicatives qui divisent le mieux les échantillons d'apprentissage initiaux (racines) en deux groupes disjoints (nœuds) en fonction de la modalité ou de la valeur. Pour ce faire, nous posons toutes les questions binaires possibles à partir des variables explicatives et choisissons la "meilleure" répartition parmi l'ensemble des répartitions autorisées testées.
Une division est dite réalisable si aucun des sous-nœuds résultants n'est vide.
Par conséquent, il est impératif de déterminer les critères selon lesquels les partitions sont considérées comme optimales. Ce critère de découpage est basé sur la définition de la fonction Hétérogénéité : Le but est de diviser les individus en deux groupes aussi homogènes que possible quant aux variables qu'ils décrivent.
L'hétérogénéité des nœuds est mesurée par une fonction non négative.
— zéro uniquement si les nœuds sont homogènes : tous les individus ont la même valeur Y
— valeur maximale lorsque les valeurs Y sont également probables ou hautement distribuées.

Pour les variables de réponse continues, cette fonction est la variance intra-nœud. Notez que N est le premier nœud et que sa séparation crée deux nœuds. Le nœud de gauche est NG et le nœud de droite est ND.
Pour tous les découpages légaux du nœud N, l'algorithme conserve celui qui maximise. Δ :

$$\Delta=\sum_{i\in n}(yi-\underset{N}{\bar{y}})^2-(\sum_{i\in NG}(yi-\underset{NG}{\bar{y}})^2+\sum_{i\in ND}(yi-\underset{ND}{\bar{y}})^2)$$

où $\underset{N}{\bar{y}}$ est la moyenne empirique de Y pour les observations associées au nœud N.
En effet, cette partition sera la partition qui classera le mieux les données et minimisera l'erreur de prédiction.
2. L'opération est répétée avec les deux nœuds nouvellement acquis. Chaque nœud correspond à un nouvel enregistrement et peut se scinder en deux parties.

3. Le processus se termine lorsque tous les nœuds créés sont homogènes. H Si aucune autre division n'est autorisée (c'est-à-dire qu'il y a des nœuds non vides) ou si le nombre d'observations incluses est inférieur au seuil, Cela évite les découpages inutiles. L'induction donne donc l'arbre maximal à plusieurs feuilles.
4. Ensuite, attribuez une valeur prédite de Y à chaque feuille. Il s'agit de la moyenne empirique des observations Y (pour les variables quantitatives) pour les individus dans les feuilles.
Ainsi, à chaque itération, l'algorithme CART a réordonné les données autant que possible. Cependant, cela ne suffit pas pour bien prédire la variable de réponse Y. Comme expliqué ci-dessus, les arbres couvrants ne sont pas les meilleurs estimateurs car ils ont trop de variance (erreur due à la variation des données). Par conséquent, la taille de l'arbre doit être optimisée.

Phase 2 : Elagage de l'arbre

L'arbre couvrant construit à l'étape précédente est très sophistiqué et Une estimation très précise du modèle d'entraînement utilisé. D'autre part, le modèle résultant est soumis au risque de surajustement, car les estimations dépendent fortement des données ajustées. Par conséquent, son pouvoir prédictif est très instable dans des échantillons de validation indépendants. Par conséquent, des modèles plus économiques et plus robustes doivent être privilégiés pour faire des prévisions sur de nouvelles normes.
C'est donc le but de la technique de taille. C'est-à-dire choisir le sous-arbre du plus grand arbre qui donne des prédictions robustes sur une base de test indépendante.

Cette procédure est décrite dans l'algorithme suivant :
1. Construire une séquence imbriquée de sous-arbres de l'arbre maximal
Soit T_{max} l'arbre maximum et K le nombre de feuilles. K représente la complexité de l'arbre.

En théorie, la mesure de qualité d'un arbre T est l'erreur $E[(Y - T(X)^2]$. Ne pouvant pas la calculer, on se limite à l'estimer par l'erreur empirique, appelée taux d'erreur de l'arbre. Elle est définie ainsi :

$R(T) = 1/n \sum_{i=1}^{n} (Yi - T(Xi))^2$

Elle mesure, pour chaque élément i de la base, l'écart entre la valeur T(Xi) prédite par le modéle et la valeur "réelle " Yi observée dans la base. En regroupant les termes par feuilles, on peut réécrire R(T) comme suit :

$R(T) = \sum_{t \in k} R(t)$

où K désigne l'ensemble des feuilles de l'arbre T.
On serait tent´es de prendre $\hat{T} = \underset{T}{argmin} \preceq T_{max}$ R(T) comme critére de sélection de l'arbre optimal. Ce choix ménerait néanmoins à retenir T_{max}, que l'on a prouvé non optimal en termes d'efficacité et de fiabilité. Ainsi, la construction de la séquence d'arbres emboités repose sur une pénalisation de la complexité de l'arbre :

$C(T) = R(T) + \lambda K$

En effet, le deuxiéme terme est d'autant plus élevé que le nombre de feuilles K de l'arbre T considéré est grand. Le facteur λ, qui est une constante à régler, sert à doser le poids accordé à cette pénalisation par rapport au premier terme.
Pour $\lambda = 0$, $T_{max} = T_K$ minimise C(T). En faisant croitre λ, l'une des divisions de T_{max}, celle pour laquelle l'amélioration de R est la plus faible (inférieure à λ), apparait comme superflue et les deux feuilles qui en découlent sont regroupées dans le nœud pére qui devient terminal. T_K devient donc T_{K-1}.

Ce processus est répété pour construire des séquences imbriquées :

$T_{max} = T_K \supset T_{K-1} \supset ..... \supset T_1$

où $\underset{1}{T}$, le nœud racine, regroupe l'ensemble de l'échantillon.

On notera $\underset{\lambda}{\hat{T}}$ la suite d'arbres obtenus suite à l'élagage par les différentes valeurs de λ.

2. Recherche de l'arbre optimal

A ce stade, on connait une famille $\underset{\lambda}{\hat{T}}$ de sous-arbres de $\underset{max}{T}$, bons candidats pour répondre aux objectifs. Il faut à présent déterminer, parmi ce premier choix d'arbre, lequel est le meilleur prédicteur. Cela revient à choisir la juste valeur de λ. On souhaite alors évaluer :
$\mathrm{E}[(Y-\underset{\lambda}{\hat{T}}(x))^2]$.

Cette fois encore, l'estimateur le plus intuitif serait le critére empirique :

$$\hat{\lambda}=\underset{\lambda}{argmin}\mathrm{R}(\underset{\lambda}{\hat{T}})$$

mais cela nous amènerait à choisir $\hat{\lambda} = 0$, ce qui donne de nouveau $\underset{max}{T}$. Cela est du au fait que le taux d'erreur est estimé par re-substitution sur l'échantillon d'apprentissage et est donc biaisé.
Une estimation sans biais est obtenue par l'utilisation d'un autre échantillon (base de validation) :

$$\hat{\lambda}= \underset{\lambda}{argmin} \underset{validation}{R} (\underset{\lambda}{\hat{T}})$$

Finalement, à la valeur de λ minimisant l'estimation sans biais de l'erreur de prévision correspond l'arbre jugé optimal. [12]

2.0.2 Principe de l'algorithme CART :

*L'algorithme commence avec un nœud racine contenant l'ensemble complet des données d'entraînement.
*Il sélectionne la meilleure variable d'entrée pour diviser les données en deux sous-ensembles (nœuds fils) en utilisant une mesure de qualité de partitionnement (par exemple, l'indice de Gini pour la classification ou l'erreur quadratique moyenne pour la régression).

*L'algorithme récursivement répète cette étape pour chaque nœud fils jusqu'à ce qu'une condition d'arrêt soit atteinte. Les conditions d'arrêt peuvent être basées sur des critères tels que la profondeur maximale de l'arbre, le nombre minimum d'échantillons par nœud ou d'autres critères de pureté.
*Une fois l'arbre construit, les prédictions sont effectuées en parcourant l'arbre à partir du nœud racine jusqu'à une feuille correspondante à un sous-ensemble spécifique des données. La valeur de sortie associée à cette feuille est utilisée comme prédiction. [12]

2.0.3 Hypothèse de l'algorithme CART :

L'hypothèse sous-jacente de l'algorithme CART est que les variables d'entrée sont divisibles en sous-groupes homogènes (pour la classification) ou qu'elles ont des relations linéaires avec la variable cible (pour la régression). Cela signifie que l'algorithme suppose qu'il existe des seuils ou des combinaisons de seuils dans les variables d'entrée qui permettent de séparer les classes ou de modéliser la relation de manière optimale.
L'algorithme CART est populaire en raison de sa simplicité, de sa facilité d'interprétation et de sa capacité à gérer des données mixtes (variables continues et catégorielles). Cependant, il peut être sensible aux variations des données d'entraînement et peut être sujet à un surajustement (overfitting) si l'arbre est trop profond ou complexe. Des techniques comme la taille maximale de l'arbre, la sélection de variables et la validation croisée peuvent être utilisées pour atténuer ces problèmes potentiels. [12]

2.0.4 Création des groupes de véhicules par CART :

La création de groupes de véhicules est une tâche importante pour de nombreuses entreprises impliquées dans la vente ou la location de véhicules. Cette tâche peut être facilitée par l'utilisation de l'algorithme CART, qui permet de créer des groupes de véhicules en fonction de leurs caractéristiques communes.

L'algorithme CART est un algorithme de classification d'arbres de décision qui peut être utilisé pour créer un arbre de décision à partir d'un ensemble de données. L'arbre de décision est un modèle qui peut être utilisé pour classer de nouvelles données en fonction de leurs caractéristiques.

En utilisant l'algorithme CART pour créer des groupes de véhicules, nous pouvons classifier les véhicules en fonction de leurs caractéristiques telles que la marque, le modèle, l'année de fabrication, le kilométrage et le type de carburant. Une fois que l'arbre de décision a été créé, il peut être utilisé pour prédire le groupe de véhicule auquel appartient un nouveau véhicule.

Dans ce processus, les données de véhicules sont généralement collectées sous forme de tableaux. Dans le code Python, nous pouvons utiliser la bibliothèque scikit-learn pour créer l'arbre CART à partir de ces tableaux. La bibliothèque permet également de diviser les données en ensembles d'entraînement et de test, d'encoder les variables catégorielles et de calculer la précision du modèle.

Une fois que l'arbre CART est créé et validé, il peut être utilisé pour créer des groupes de véhicules en fonction de leurs caractéristiques communes. Cela permet aux entreprises impliquées dans la vente ou la location de véhicules de mieux comprendre leurs clients et de leur proposer des offres personnalisées en fonction de leurs besoins. [12]

Troisième partie

partie Pratique :

2.1 Présentation des données :

Les données de notre étude proviennent du prime d'assurance automobile de la compagnie d'assurance SAA portant de deux tableaux ('sinistre','production')sur cinq ans,la description de ces données se fait sur deux types de variables,variables qualitatives et variables quantitatives.
Nous avons aussi d'autre variables dans notre table de données qui donnent une autre spécification sur le choix du client comme le ID de chaque contrat,le type véhicule de chaque client,le code agence,garantie,....etc

	code_agenc	'libe_agence	'Code_dr'	'numepoli'	'numesini'	'datedecl'	'MT_SIN_DE(	'codecate'	'codegara
2	1601	BLIDA " A"	16	1100013673	110268	07/09/2017	54000	1110	TR
3	1602	BLIDA " B"	16	1100017019	110293	02/07/2017	19000	1110	DCD
4	1602	BLIDA " B"	16	1100019448	110384	03/09/2017	7500	1110	PEA
5	1602	BLIDA " B"	16	1100010288	110501	06/11/2017	9700	1110	DCC
6	1602	BLIDA " B"	16	1100016280	110604	25/12/2017		1110	DR
7	1603	BLIDA " C "	16	1100032246	110857	03/09/2017	54000	1110	TR
8	1603	BLIDA " C "	16	1511000032	150004	03/12/2017	48000	1511	DDEC
9	1604	BOUFARIK	16	1100029961	110212	16/02/2017	44000	1110	RC
10	1604	BOUFARIK	16	1100033119	110338	21/03/2017	44000	1110	RC
11	1604	BOUFARIK	16	1100035614	110761	21/06/2017	44000	1110	RC
12	1604	BOUFARIK	16	1100039884	110948	25/07/2017	19000	1110	DCG
13	1604	BOUFARIK	16	1100037005	110979	01/08/2017	19000	1110	BDG
14	1604	BOUFARIK	16	1100037056	110980	03/08/2017	7500	1110	PEA
15	1604	BOUFARIK	16	1100046007	111104	05/09/2017	44000	1110	RC
16	1604	BOUFARIK	16	1100043396	111193	24/09/2017	38500	1110	DASC3
17	1604	BOUFARIK	16	1100005222	111255	08/10/2017	44000	1110	RC
18	1604	BOUFARIK	16	1100046652	111442	16/11/2017	19000	1110	DCD
19	1604	BOUFARIK	16	1100046742	111452	19/11/2017	7500	1110	PEA
20	1605	CHERCHELL	16	1100008133	110238	24/05/2017	10000	1110	RVF
21	1606	HADJOUT	16	1100017561	110557	09/08/2017	10000	1110	RVF
22	1607	BOU ISMAIL	16	1100011464	110015	08/01/2017	54000	1110	TR
23	1607	BOU ISMAIL	16	1100001010	110041	22/01/2017	44000	1110	RC

FIGURE 2.2 – tableau des données de sinistre

	code_agence	Nom_agence	agence_rattachement	code_DR	Nom_DR	type_agence	DATE_miseCIRCULA	NUMEIMMA	DATE_DELIV_PERMI	TYPE_PERMIS	COD_NAT_VE	ID
2	4,161	TIPAZA	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2015	08205.115.42	19/02/2019	B	1.0	1100011819-1610-5
3	5,1807	BENI SLIMANE	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2018	013836.118.16	22/02/2015	B	1.0	1100020857-1807-0
4	6,1807	BENI SLIMANE	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2018	013836.118.16	22/02/2015	B	1.0	1100020857-1807-1
5	7,171	TENIET EL HAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2011	00532.111.38	07/08/2006	B	0.0	1100004643-1710-14
6	8,171	TENIET EL HAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2011	05948.311.44	29/01/2012	B	0.0	1100007559-1710-6
7	9,1807	BENI SLIMANE	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2009	09225.309.26	25/01/1987	B	1.0	1100019071-1807-2
8	10,1708	EL KHEMIS	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2013	05945.113.44	19/02/2007	B	1.0	1100014174-1708-7
9	11,1807	BENI SLIMANE	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2008	05386.208.26	13/11/1999	B	1.0	1100015253-1807-8
10	12,1809	BIRINE	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2012	13562.112.17	19/05/2006	B	1.0	1100006929-1809-1
11	13,171	TENIET EL HAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2017	00056.317.38	28/01/2014	B	0.0	1100008836-1710-1
12	14,1811	HASSI Bahbah	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2012	05171.112.17	05/11/1994	B	1.0	1100006816-1811-9
13	15,1807	BENI SLIMANE	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2016	055350.116.16	26/05/2010	B	1.0	1100020072-1807-3
14	16,1804	MEDEA	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2016	00168.116.26	01/01/2010	B	1.0	1100020682-1804-2
15	17,171	TENIET EL HAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2010	04559.210.38	12/06/2001	B	0.0	1100008821-1710-2
16	18,171	TENIET EL HAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2015	01722.315.38	26/07/2015	B	0.0	1100008806-1710-1
17	19,1905	DJELFA	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2011	02107.111.17	10/04/2002	B	1.0	1100015410-1905-5
18	20,1708	EL KHEMIS	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2013	05945.113.44	19/02/2007	B	1.0	1100014174-1708-6
19	21,171	TENIET EL HAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2010	04559.210,38	12/06/2001	B	0.0	1100008821-1710-3
20	22,1701	MILIANA	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2005	08521.105.44	02/01/2008	B	1.0	1100009673-1701-2
21	23,1906	MESSAAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2009	04059.309.17		B	0.0	1100022458-1906-20
22	24,1906	MESSAAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/2013	06800.313.17		B	0.0	1100022458-1906-20
23	25,1906	MESSAAD	16	16	Direction RÃ©gionale MOUZAI	Agence directe	01/01/1985	02405.285.17		B	0.0	1100022458-1906-20

FIGURE 2.3 – tableau des données production

2.1.1 Nettoyage des données :

Nous avons effectué un nettoyage des données à partir de deux fichiers : ”production” et ”sinistre”. Les données couvrent une période de cinq ans, de 2017 à 2021. Chaque année est représentée par un tableau distinct. Nous avons réalisé une jointure en utilisant la colonne ”ID” comme lien entre les tableaux de chaque année en un seul tableau consolidé.

JOINTURE DE DEUX FICHIERS : Avant d'utiliser l'algorithme GLM et l'algorithme CART nous avons grouper les deux fichiers ”production” et ”Sinistre ” dans un seul fichier que nous avons nommé ”BASE”. Cette étape était nécessaire pour regrouper toutes les données pertinentes dans un seul ensemble.voici le tableau de BASE :

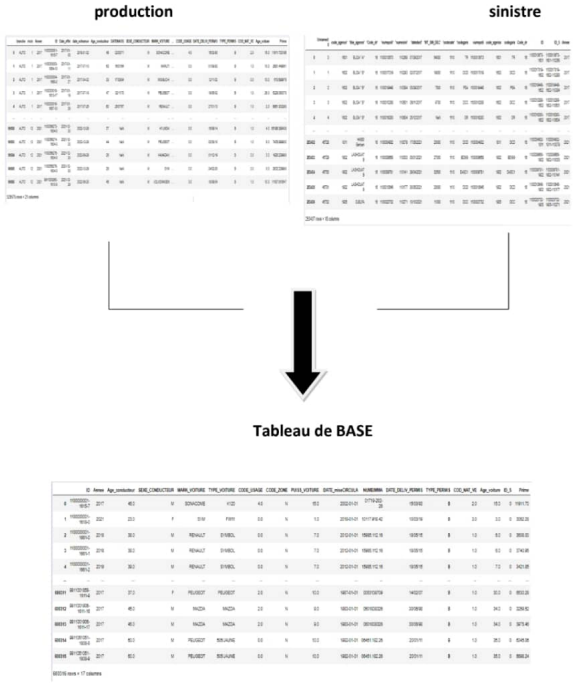


FIGURE 2.4 – tableau de BASE

Cette analyse de la base des données ”production” nous a permis d'étudier comment le sexe du conducteur était associé à son niveau d'implication dans différentes années, grâce à une visualisation claire et informative. Les résultats obtenus contribuent à une meilleure compréhension des éventuelles différences dans l'implication des conducteurs en fonction de leur sexe et de l'année étudiée.voici le diagramme :

```
Entrée [36]: # reading the database

sns.barplot(x='Annee',y='Prime', data=production,
            hue='SEXE_CONDUCTEUR')

plt.show()
```

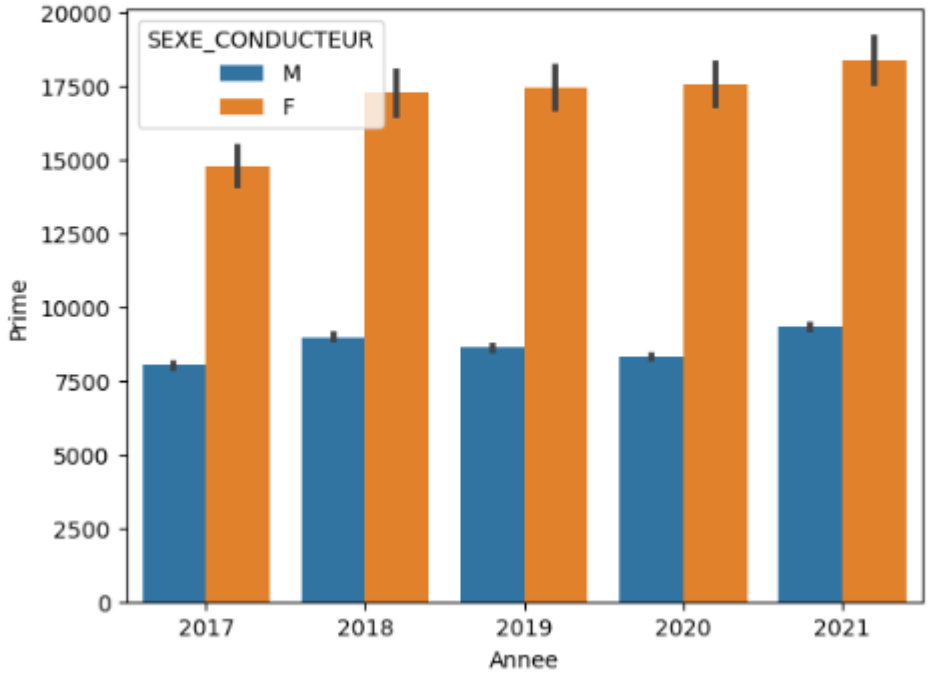


FIGURE 2.5 – graphique à barres du sexe conducteur

Une matrice de corrélation est un tableau qui représente les relations statistiques entre les variables d'un ensemble de données. Elle mesure la corrélation, c'est-à-dire la relation linéaire, entre paires de variables. Chaque cellule de la matrice contient un coefficient de corrélation qui indique la force et la direction de la relation entre deux variables,Les coefficients de corrélation peuvent varier de -1 à 1. Un coefficient de corrélation de 1 indique une corrélation positive parfaite, ce qui signifie que les variables évoluent en parfaite harmonie dans la même direction. Un coefficient de corrélation de -1 indique une corrélation négative parfaite, ce qui signifie que les variables évoluent de manière opposée. Un coefficient de corrélation de 0 indique l'absence de corrélation linéaire entre les variables.

Cette visualisation de la matrice de corrélation permet de mieux comprendre les relations entre les différentes variables de la base de données ”production”. Elle peut aider à identifier des corrélations positives, négatives ou neutres, ce qui peut être utile pour explorer les relations entre les variables et détecter des associations potentielles entre elles.voici notre matrice de correlation :

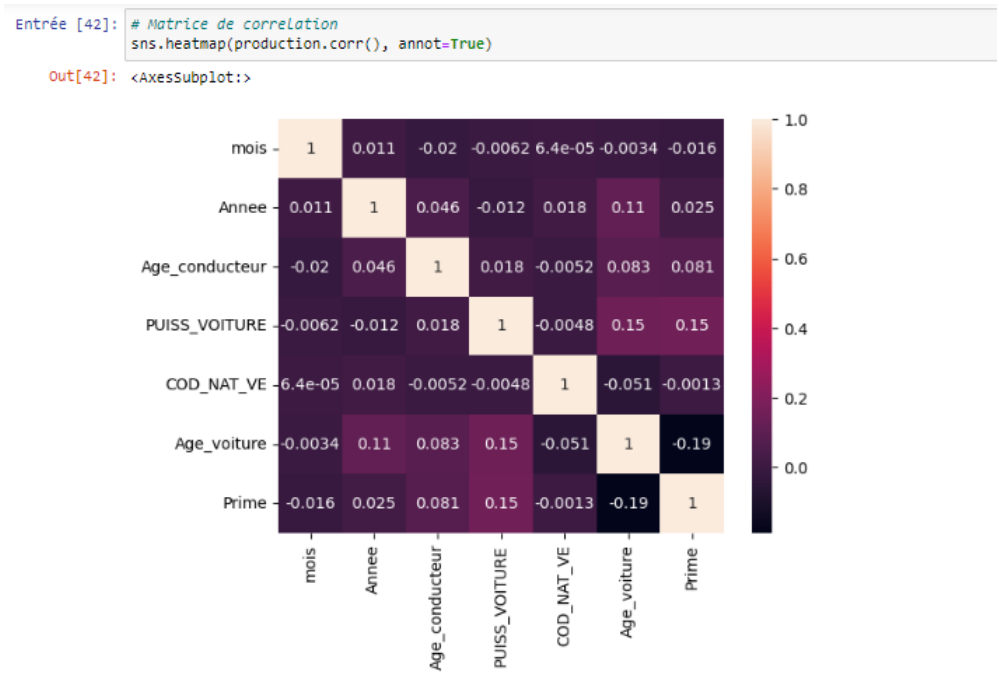


FIGURE 2.6 – Matrice de correlation

Dans cette matrice de corrélation, nous comparons la variable ’Prime’ avec ’Age conducteur’, ’PUISS VOITURE’ et ’Age voiture’. Les résultats montrent des corrélations intéressantes. Nous observons une corrélation négative (-0.19) entre ’Prime’ et ’Age voiture’, ce qui suggère qu’une augmentation de l’âge du véhicule est associée à une baisse de la prime d’assurance. De plus, nous constatons une corrélation positive (0.15) entre ’Prime’ et ’PUISS VOITURE’, indiquant qu’une augmentation de la puissance du véhicule est liée à une augmentation de la prime d’assurance. Enfin, nous relevons également une corrélation positive (0.081) entre ’Prime’ et ’Age conducteur’, ce qui suggère qu’une augmentation de l’âge du conducteur est associée à une augmentation de la prime d’assurance.

2.2 Application des méthodes :

Avant d'utiliser des méthodes de machine learning, il est important de réaliser une étape préliminaire de transformation des données. Cette étape vise à convertir les colonnes catégoriques en données numériques permettant ainsi de préparer le fichier source dans un format approprié pour les algorithmes GLM et CART.
Pour l'algorithme GLM, la conversion en format numérique permet d'appliquer des méthodes statistiques basées sur des équations linéaires. Les variables catégorielles sont transformées en variables indicatrices ou en variables binaires, ce qui permet de les incorporer correctement dans le modèle GLM.Pour l'algorithme CART, la conversion des données en format numérique est nécessaire pour construire un arbre de décision. Les variables catégorielles sont généralement encodées sous forme de variables binaires, où chaque catégorie est représentée par une colonne distincte.
En convertissant les variables catégorielles en valeurs numériques pour les algorithmes GLM et CART, nous avons pu préparer les données de manière appropriée et les rendre compatibles avec ces méthodes d'analyse. Cela nos permis d'exploiter pleinement les avantages de ces algorithmes dans mon étude et d'obtenir des résultats significatifs pour mes analyses et mes modèles.

2.2.1 Modélisation avec GLM :

L'algorithme GLM est une méthode d'analyse statistique qui permet de modéliser les relations entre des variables dépendantes et des variables indépendantes dans un cadre général.Nous avons utilisé cet algorithme pour modéliser les relations entre mes variables explicatives et ma variable cible. Dans mes fichiers, la variable cible était la "Prime". Nous avons choisi d'utiliser une distribution Gamma pour modéliser cette variable en fonction du type de données que n'avais.
En utilisant l'algorithme GLM avec une distribution Gamma et en appliquant une régression logistique, nous avons pu modéliser les relations entre mes variables explicatives et ma variable cible "Prime" dans mes données. Cette approche offre une flexibilité pour modéliser des variables cibles qui suivent différentes distributions de probabilité, en fonction de la nature des données.

voici une Régression logistique de la loi Gamma :

Generalized Linear Model Regression Results

Dep. Variable:	Prime	No. Observations:	779512
Model:	GLM	Df Residuals:	779507
Model Family:	Gamma	Df Model:	4
Link Function:	inverse_power	Scale:	4.4993
Method:	IRLS	Log-Likelihood:	nan
Date:	Sat, 15 Jul 2023	Deviance:	2.0187e+07
Time:	22:56:37	Pearson chi2:	3.51e+06
No. Iterations:	100	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0101	2.88e-05	350.998	0.000	0.010	0.010
Age_conducteur	-8.154e-06	1.96e-08	-416.148	0.000	-8.19e-06	-8.12e-06
Annee	-4.732e-06	1.37e-08	-345.013	0.000	-4.76e-06	-4.7e-06
PUISS_VOITURE	-1.004e-05	2.41e-08	-415.992	0.000	-1.01e-05	-1e-05
Age_voiture	9.014e-06	2.17e-08	415.186	0.000	8.97e-06	9.06e-06

BIC: 9612359.649661321

FIGURE 2.7 – Régression logistique par la loi Gamma

'z' : le score z est une mesure de la déviation d'une variable par rapport à sa moyenne. Dans le contexte du modèle GLM, le score z est utilisé pour évaluer la significativité statistique des coefficients des variables explicatives. Un score z élevé indique que le coefficient est significativement différent de zéro, ce qui suggère une influence significative de la variable sur la variable cible.
'P' : la valeur P (p-value) est une mesure de la probabilité d'obtenir des résultats aussi extrêmes que ceux observés, sous l'hypothèse nulle selon laquelle le coefficient de la variable explicative est égal à zéro. Une valeur P faible (généralement inférieure à 0,05) indique une significativité statistique, ce qui signifie qu'il y a des preuves solides que la variable explicative a un impact significatif sur la variable cible.
'std err' : il s'agit de l'erreur standard, qui est une mesure de la précision de l'estimation du coefficient de la variable explicative. Une erreur standard plus faible indique une estimation plus précise et plus fiable du coefficient.
Nous avons effectué un test, le test du BIC, afin de déterminer quelle est la meilleure loi entre la loi Gamma et la loi gaussienne.
BIC : Le BIC (Bayesian Information Criterion), également connu sous le nom de

critère d'information bayésien, est une mesure statistique utilisée pour évaluer la qualité d'un modèle statistique. Il est largement utilisé pour la sélection de modèle et la comparaison de modèles alternatifs, BIC est calculé à partir de la fonction de vraisemblance du modèle et du nombre de paramètres du modèle. Il est donné par l'expression BIC = -2 $\times$ logvraisemblance + k $\times$ log(n), où logvraisemblance est la log-vraisemblance maximisée du modèle, k est le nombre de paramètres du modèle et n est la taille de l'échantillon de données.

une éqaution de la Régression logistique d'une *loi Gamma* :

$\hat{y}$ = 0.0101 -8.154e-06 $\times Age\_conducteur - 4.732e-06 \times Annee - 1.004e-05 \times PUISS\_VOITURE + 9.014e-06 \times Age\_voiture$

Le graphe tracé représente les valeurs prédites par rapport aux valeurs réelles de la variable cible :

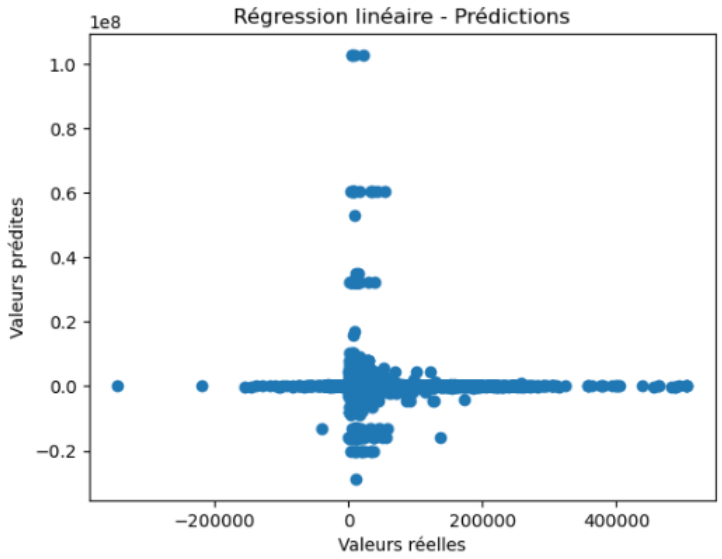


FIGURE 2.8 – graphe de la régression linéaire - prédictions

Le graphique de dispersion montre la relation entre les valeurs réelles et les valeurs prédites de la variable cible, dans le contexte d'une régression linéaire généralisée. Les valeurs réelles sont représentées sur l'axe des abscisses (x) et les valeurs prédites sont représentées sur l'axe des ordonnées (y).
Si les valeurs prédites correspondent étroitement aux valeurs réelles, les points

sur le graphique seront alignés approximativement sur une ligne droite. Cela indiquerait une bonne capacité du modèle à prédire la variable cible.En revanche, si les valeurs prédites diffèrent considérablement des valeurs réelles, les points sur le graphique seront dispersés de manière irrégulière. Cela peut indiquer des problèmes dans la modélisation ou des variables manquantes qui ne sont pas prises en compte par le modèle En examinant le graphique, il est important de rechercher une tendance générale. Si les points sont répartis de manière cohérente autour d'une ligne diagonale ascendante, cela suggère une bonne adéquation du modèle. Si les points sont dispersés de manière désordonnée sans motif apparent, cela peut indiquer des problèmes dans le modèle ou la nécessité d'explorer d'autres variables explicatives.

Le graphe représente un Q-Q plot des résidus d'un modèle de régression linéaire généralisée avec une famille de distribution gamma :

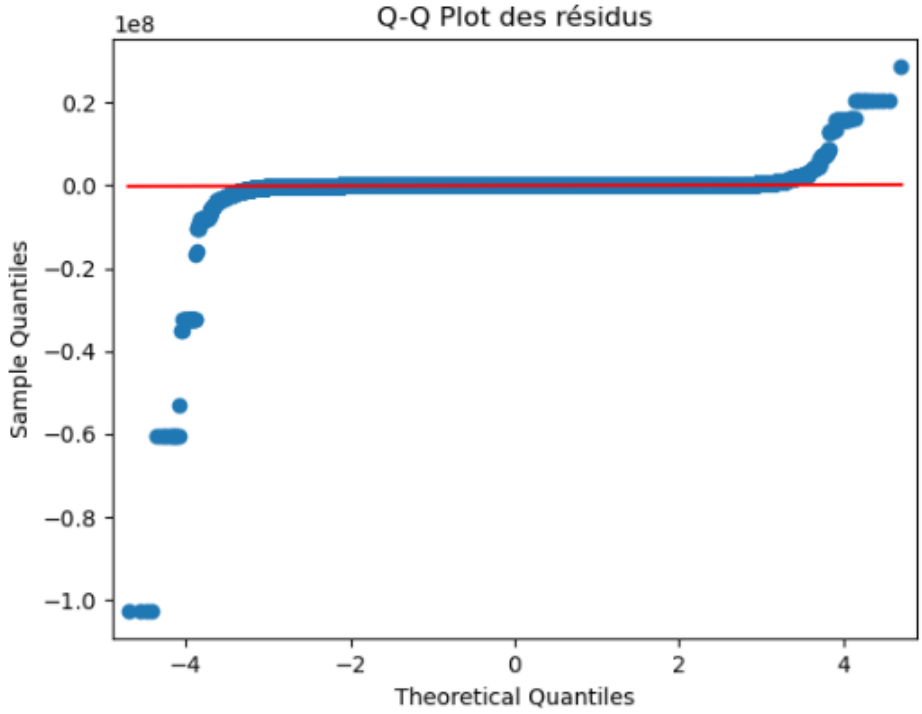


FIGURE 2.9 – Q-Q plot des résidus

Le Q-Q plot (Quantile-Quantile plot) est un graphique utilisé pour évaluer si les résidus d'un modèle statistique suivent approximativement une distribution théorique spécifiée. Dans ce cas, le Q-Q plot est utilisé pour vérifier si les résidus du modèle de régression linéaire généralisée suivent une distribution gamma.
Le graphique compare les quantiles théoriques d'une distribution (ici, la distribution gamma) aux quantiles empiriques des résidus du modèle. Les quantiles sont essentiellement les valeurs correspondantes à des probabilités spécifiques. Si les résidus suivent la distribution théorique, les points sur le graphique devraient approximativement suivre une ligne droite, indiquant une bonne adéquation du modèle.Dans le Q-Q plot, les résidus sont représentés sur l'axe des ordonnées (y) et les quantiles théoriques de la distribution gamma sont représentés sur l'axe des abscisses (x). Une ligne rouge (ici, line="r") est tracée pour représenter la ligne idéale si les résidus suivent exactement la distribution théorique.
L'interprétation du Q-Q plot se fait en examinant la proximité des points par rapport à la ligne idéale. Si les points sont alignés approximativement le long de la ligne rouge, cela indique une bonne adéquation du modèle aux résidus. Si les points s'éloignent considérablement de la ligne rouge, cela peut indiquer des problèmes dans le modèle ou des violations des hypothèses de la distribution gamma.

voici une Régression logistique de la loi gaussienne :

Generalized Linear Model Regression Results

Dep. Variable:	Prime	No. Observations:	779512
Model:	GLM	Df Residuals:	779507
Model Family:	Gaussian	Df Model:	4
Link Function:	identity	Scale:	1.0562e+08
Method:	IRLS	Log-Likelihood:	-8.3070e+06
Date:	Tue, 27 Jun 2023	Deviance:	8.2333e+13
Time:	15:09:18	Pearson chi2:	8.23e+13
No. Iterations:	3	Pseudo R-squ. (CS):	0.2123
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.415e+05	1.7e+04	-31.786	0.000	-5.75e+05	-5.08e+05
Age_conducteur	153.0961	0.667	229.510	0.000	151.789	154.403
Annee	268.4413	8.438	31.813	0.000	251.903	284.980
PUISS_VOITURE	678.4342	3.332	203.605	0.000	671.903	684.965
Age_voiture	-367.4528	1.882	-195.279	0.000	-371.141	-363.765

BIC: 82332739007644.03

FIGURE 2.10 – Régression logistique de la loi gaussienne

une éqaution de la Régression logistique d'une *loi gaussienne* :

$\hat{y}$= -5.415e+05 +153.0961 $\times Age\_conducteur + 268.4413 \times Annee + 678.4342 \times PUISS\_VOITURE - 367.4528 \times Age\_voiture$

Le graphe représente un diagramme de dispersion des résidus pour un modèle de régression linéaire :

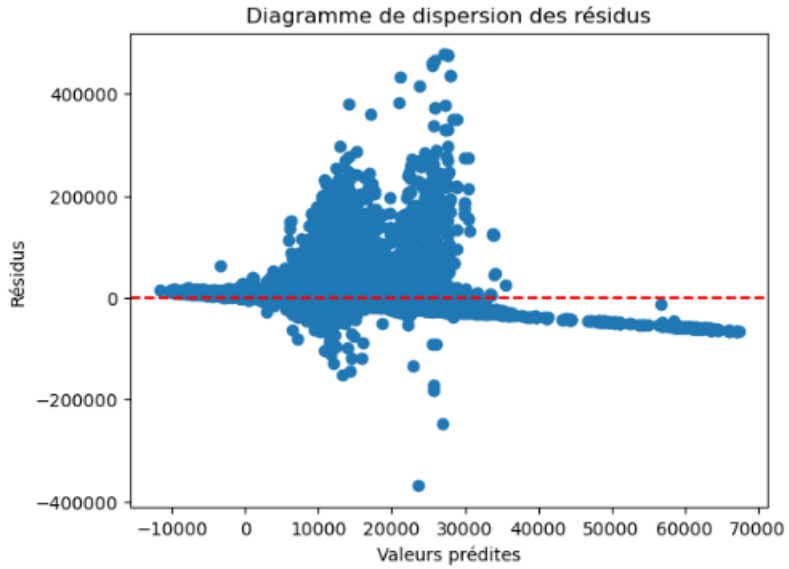


FIGURE 2.11 – Régression logistique de la loi gaussienne

Le diagramme de dispersion des résidus est un outil graphique utilisé pour évaluer la qualité d'ajustement d'un modèle de régression. Il représente la relation entre les valeurs prédites par le modèle (sur l'axe des x) et les résidus du modèle (sur l'axe des y).
les résidus sont calculés en soustrayant les valeurs prédites par le modèle des valeurs observées. Les résidus représentent donc les erreurs du modèle, c'est-à-dire la différence entre les valeurs réelles et celles prédites par le modèle, Idéalement, on voudrait que les résidus soient dispersés de manière aléatoire autour de zéro, sans aucune tendance ou relation systématique avec les valeurs prédites,Si les résidus sont répartis de manière aléatoire autour de zéro, cela indique que le modèle est capable de capturer la variation dans les données et que les erreurs sont bien distribuées. Cela renforce la validité de l'ajustement du modèle et permet d'utiliser les estimations des coefficients et les tests statistiques appropriés.

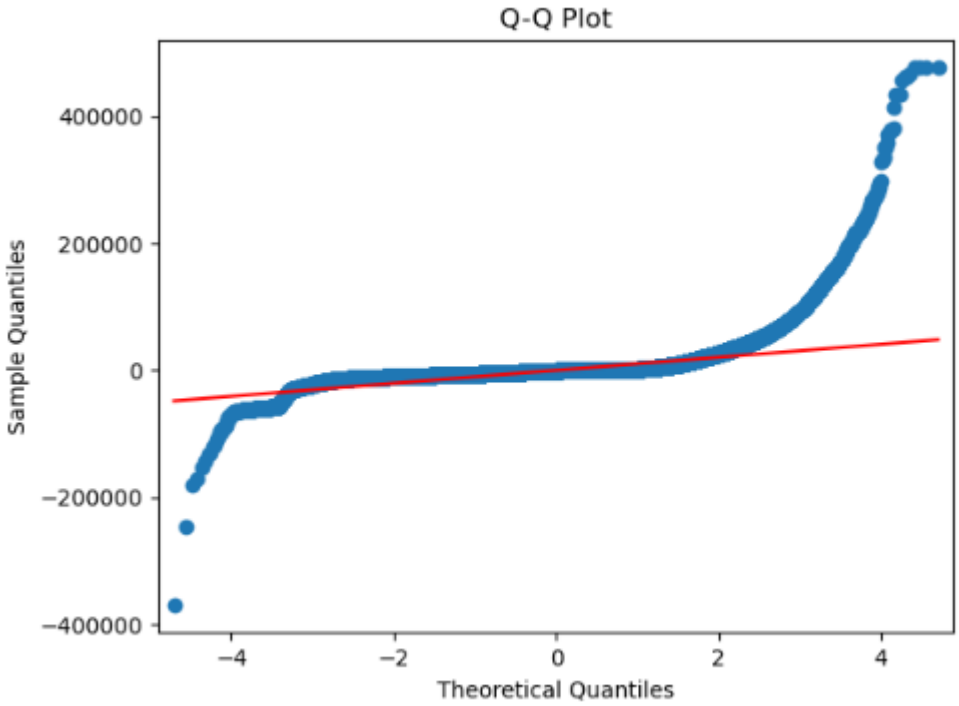


FIGURE 2.12 – Q_Q plot

les résidus du modèle sont calculés et un Q-Q plot (Quantile-Quantile plot) est tracé pour évaluer la distribution des résidus. Le Q-Q plot compare les quantiles des résidus observés avec les quantiles d'une distribution théorique (ici, une distribution gaussienne). Si les résidus suivent approximativement une distribution gaussienne, les points sur le graphique suivront approximativement une ligne droite.
Si les points du Q-Q plot se rapprochent de la ligne droite, cela suggère que les résidus suivent une distribution gaussienne et que le modèle est approprié. Cependant, si les points s'écartent considérablement de la ligne droite, cela indique une violation de l'hypothèse de distribution gaussienne des résidus et nécessite une réévaluation du modèle.

2.2.2 Traitement des données CART :

Le traitement des données d'algorithme CART sur Python consiste à préparer les données pour être utilisées dans l'algorithme CART. Voici les étapes principales pour traiter les données :

Préparation des données : Tout d'abord, nous avons préparer nos données en les divisant en un ensemble d'entraînement et un ensemble de test. Nous avonségalement effectuer les étapes de nettoyage des données, telles que le traitement des valeurs manquantes, l'encodage des variables catégorielles, et la normalisation ou la mise à l'échelle des variables numériques si nécessaire.

Importation des bibliothèques :Nous devons importer les bibliothèques nécessaires dans notre script Python, notamment scikit learn (sklearn) qui offre des outils pour l'implémentation de l'algorithme CART.

Création du modèle CART : Nous pouvons créer un modèle CART en utilisant la classe DecisionTreeClassifier de scikit learn pour la classification ou la classe DecisionTreeRegressor pour la régression. Nous pouvons spécifier des paramètres tels que la profondeur maximale de l'arbre, le critère de partitionnement (par exemple, l'indice de Gini ou l'entropie), etc.

Entraînement du modèle : Nous devons entraîner notre modèle CART en utilisant les données d'entraînement. Utilisez la méthode fit(X, y) en passant les features (X) et les étiquettes (y) correspondantes.

Prédiction : Une fois que notre modèle est entraîné, nous pouvons l'utiliser pour effectuer des prédictions sur de nouvelles données. Utilisez la méthode predict(X test) en passant les données de test (X test) pour obtenir les prédictions.

Évaluation du modèle : Évaluez les performances de notre modèle en comparant les prédictions avec les valeurs réelles. Selon le problème, nous pouvons utiliser des métriques telles que l'exactitude (accuracy), la précision (precision), le rappel (recall), le score F1 (F1 score) ou d'autres métriques appropriées.

Optimisation du modèle : nous pouvons ajuster les hyperparamètres de notre modèle CART pour obtenir de meilleures performances. Cela peut inclure l'utilisation de la validation croisée, du réglage des hyperparamètres avec une recherche

par grille (grid search), ou d'autres techniques d'optimisation.

Ces étapes générales nous donnent une idée de la façon dont les données sont traitées lors de l'utilisation de l'algorithme CART en Python. Cependant, il est important de noter que les détails précis du traitement des données peuvent varier en fonction de notre ensemble de données spécifique et des exigences de notre problème.

Paramétrage et implémentation de l'algorithme CART :

L'algorithme CART (Classification and Regression Trees) est un algorithme utilisé pour la construction d'arbres de décision binaires. Il peut être utilisé pour la classification ou la régression en fonction du type de variable cible. Voici comment nous pouvons configurer et implémenter l'algorithme CART en utilisant Python :
Étape 1 : Préparation des données : Avant d'implémenter l'algorithme CART, nous devons préparer nos données en effectuant les étapes suivantes :
*Importez les bibliothèques nécessaires, telles que numpy et pandas, pour manipuler les données.
*Chargez nos données dans une structure de données appropriée, comme un DataFrame pandas.
*Séparez nos données en variables d'entrée (X) et variable cible (y).
Étape 2 : Implémentation de l'algorithme CART : Voici les étapes pour implémenter l'algorithme CART :
*Définissez une fonction pour calculer la mesure de l'impureté des nœuds (par exemple, l'indice de Gini ou l'entropie).
*Définissez une fonction récursive pour construire l'arbre de décision. Cette fonction prendra en compte les paramètres suivants : les données d'entrée (X et y), l'impureté du nœud parent, le critère d'arrêt (par exemple, la profondeur maximale de l'arbre ou le nombre minimum d'échantillons requis pour scinder un nœud), et d'autres paramètres pertinents.
*À chaque étape de la construction de l'arbre, choisissez la variable et la valeur de séparation qui minimisent l'impureté des nœuds fils.
*Divisez les données en fonction de la variable et de la valeur de séparation choisies.
*Appel récursif de la fonction de construction de l'arbre sur les nœuds fils jusqu'à ce que le critère d'arrêt soit atteint.

*Retournez l'arbre de décision construit.
voici l'algorithme de Table des coûts de complexité pour le modéle de coût moyen :

```
import pandas as pd
import numpy as np

# Chargement des données
production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')

# Calcul de la somme des primes par année
costs = production.groupby('Annee')['Prime'].sum()

# Création de la table des coûts de complexité
complexity = np.arange(1, len(costs) + 1)  # Complexité croissante (par exemple, années consécutives)
cost_table = pd.DataFrame({'Complexity': complexity, 'Cost': costs})

# Affichage de la table des coûts de complexité
print(cost_table)
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_4464\1776480894.py:5: DtypeWarning: Columns (8) have mixed types. Specify dtype opti
on on import or set low_memory=False.
  production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')
```

Annee	Complexity	Cost
2017	1	1.019342e+09
2018	2	9.937574e+08
2019	3	9.685965e+08
2020	4	8.113844e+08
2021	5	8.632927e+08

TABLE 2.1 – tableau des coûts de complexité pour le modéle de coût moyen

Cost : c'est le coût désigne le montant d'argent peut être mesuré de différentes maniéres,en fonction des objectifs spécifique d'une entrprise.
voici l'algorithme et la Représentation graphique de ce tableaux :

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Chargement des données
production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')

# Calcul de la somme des primes par année
costs = production.groupby('Annee')['Prime'].sum()

# Création de la table des coûts de complexité
complexity = np.arange(1, len(costs) + 1)  # Complexité croissante (par exemple, années consécutives)
cost_table = pd.DataFrame({'Complexity': complexity, 'Cost': costs})

# Affichage du graphe des coûts de complexité
plt.plot(cost_table['Complexity'], cost_table['Cost'])
plt.xlabel('Complexity')
plt.ylabel('Cost')
plt.title('Cost Complexity Graph')
plt.show()
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_4464\4124674890.py:6: DtypeWarning: Columns (8) have mixed types. Specify dtype opti
on on import or set low_memory=False.
  production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')
```

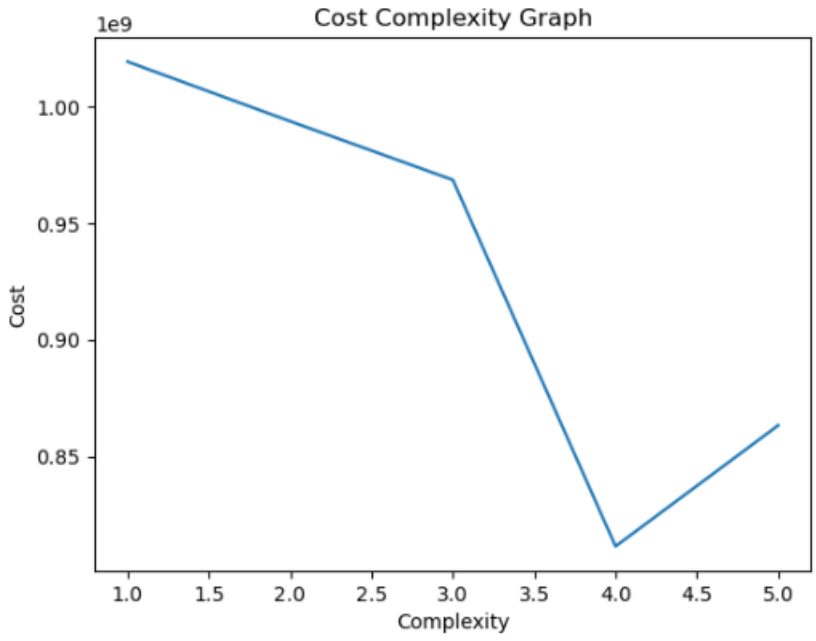


FIGURE 2.13 – Représentation graphique des coûts de complexité pour le modéle de coût moyen

Ce graphe représente les coûts de complexité pour le modèle de coût moyen On remarque que ce graphique montre une relation plus complexe entre la com-

plexité et le coût . Dans ce cas, d'autres facteurs peuvent influencer les coûts de manière non linéaire, et une analyse plus approfondie peut être nécessaire pour comprendre les tendances et les motifs spécifiques.

Mesure du pouvoir discriminant des arbres construits :

La mesure du pouvoir discriminant des arbres construits en Python fait référence à une évaluation de la capacité d'un modèle d'arbre de décision à différencier ou à discriminer entre différentes classes ou catégories dans un ensemble de données. Il existe plusieurs métriques et méthodes pour mesurer cette capacité, dont voici quelques exemples :

*Importance des variables : Cette mesure attribue un score à chaque variable d'entrée en fonction de son impact sur la pureté des nœuds de l'arbre. Une variable avec un score d'importance élevé est considérée comme ayant un pouvoir discriminant plus fort.

*Précision globale : C'est la mesure de la proportion d'observations correctement classées par l'arbre. Une précision globale élevée indique un pouvoir discriminant fort.

*Matrice de confusion : Elle présente les résultats des prédictions de l'arbre en les comparant aux véritables classes de l'ensemble de données. Une matrice de confusion clairement diagonalisée indique un pouvoir discriminant élevé, où les vrais positifs et les vrais négatifs sont correctement identifiés.

*Courbe ROC (Receiver Operating Characteristic) : Elle représente graphiquement la performance du modèle en traçant le taux de vrais positifs (Sensibilité) par rapport au taux de faux positifs (1 - Spécificité) à différents seuils de classification. Une courbe ROC s'éloignant de la ligne diagonale et se rapprochant de l'angle supérieur gauche indique un pouvoir discriminant fort.

*AUC (Area Under the Curve) : C'est une mesure numérique dérivée de la courbe ROC qui quantifie la capacité de l'arbre à discriminer entre les classes. Une valeur d'AUC proche de 1 indique un pouvoir discriminant élevé, tandis qu'une valeur proche de 0,5 indique un pouvoir discriminant faible (équivalent à une prédiction aléatoire).

Ces mesures peuvent être calculées à l'aide de bibliothèques Python telles que scikit-learn, qui offre des fonctions spécifiques pour l'évaluation des modèles d'arbre de décision.
Pour le modéle de coût moyen, nous obtenons les courbes de Lorenz suivantes.

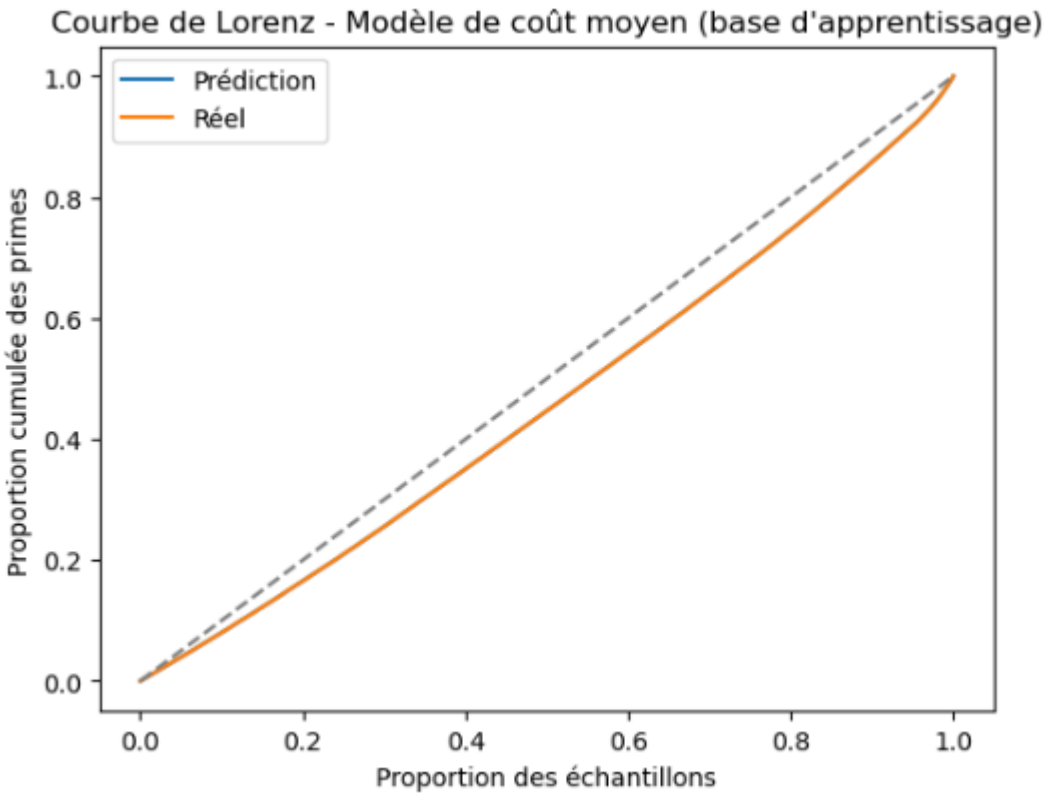


FIGURE 2.14 – Courbes de Lorenz pour le modéle de coût moyen appliqué à la base d'apprentissage

Nous affichons ci-dessous les résultats obtenus sur la base de test :

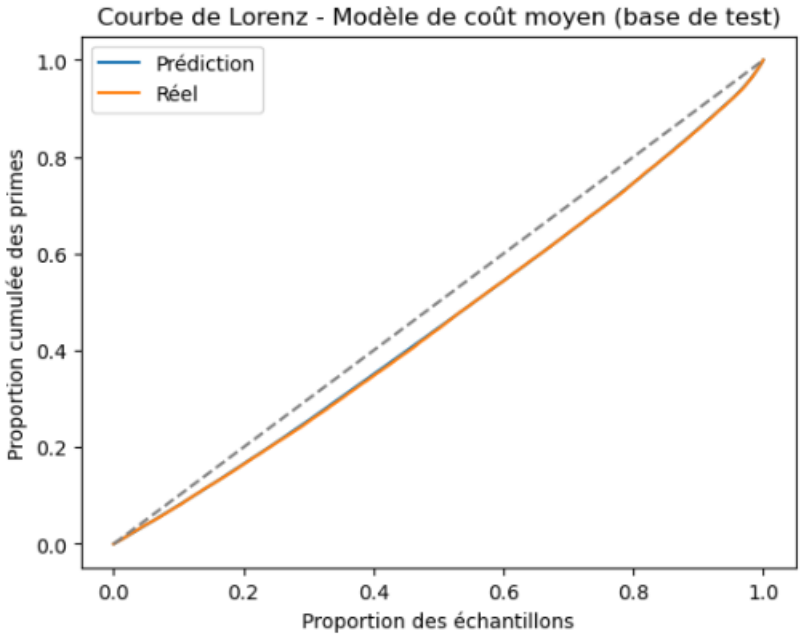


FIGURE 2.15 – Courbes de Lorenz pour le modéle de coût moyen appliqué à la base test

La courbe de Lorenz est utilisée pour évaluer la répartition cumulative des primes prédites par le modèle par rapport aux primes réelles. Elle compare les proportions cumulées des primes prédites par rapport aux proportions cumulées des primes réelles.
Dans ce cas, le modèle CART est utilisé pour prédire les primes en fonction de deux variables d'entrée : le sexe du conducteur et l'âge du conducteur. Les données d'entraînement sont divisées en ensembles d'entraînement et de test, où 80% des données sont utilisées pour l'entraînement et 20% pour les tests.
Les deux courbe de Lorenz représente le modèle de coût moyen (base d'apprentissage) de la figure 14 et le modèle de coût moyen(base de test) de la figure 15 on voix que on a la même courbe de Lorenz,La ligne en pointillés gris représente la ligne d'égalité parfaite, où les proportions cumulées prédites correspondent parfaitement aux proportions cumulées réelles,plus la courbe de prédiction se rapproche de la ligne d'égalité parfaite, meilleure est la performance du modèle. Si la courbe de prédiction se situe en dessous de la ligne d'égalité, cela indique une sous-estimation des primes, tandis que si elle se situe au-dessus, cela indique une

surestimation. Une courbe qui se rapproche de l'angle supérieur gauche (coin supérieur gauche) indique une meilleure performance du modèle, où les primes sont correctement prédites avec une plus petite proportion des échantillons.

Analysons à présent les résultats pour les arbres de fréquence.
La courbe de Lorenz est utilisée pour évaluer la répartition cumulative des fréquences de classes prédites par rapport aux fréquences réelles. Elle compare les proportions cumulées des fréquences de classes prédites par rapport aux proportions cumulées des fréquences réelles.

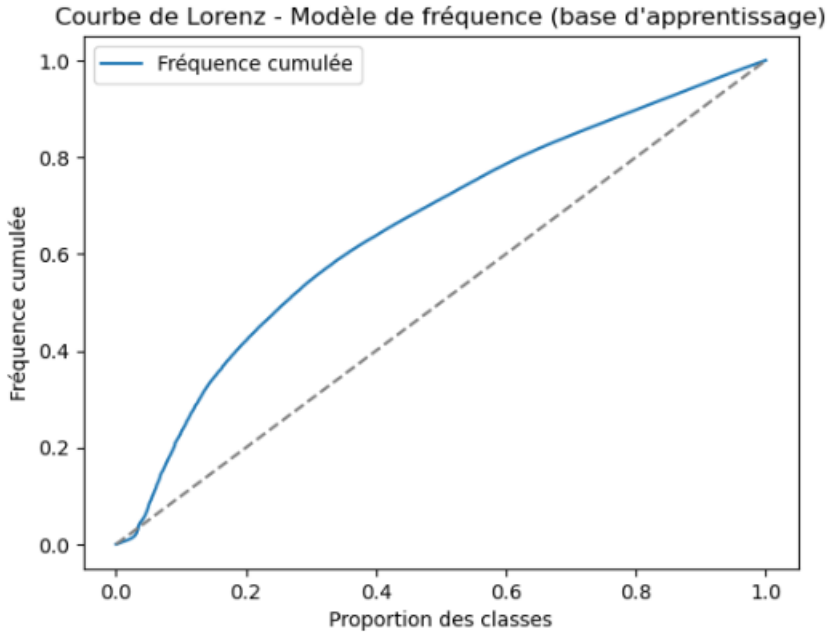


FIGURE 2.16 – Courbes de Lorenz pour le modéle de fréquence appliqué à la base d'apprentissage

cette courbe de Lorenz représente le modèle de fréquence (base d'apprentissage),Les données sont divisées en ensembles d'entraînement et de test, où 80% des données sont utilisées pour l'entraînement et 20% pour les tests,La ligne en pointillés gris représente la ligne d'égalité parfaite, où les proportions cumulées prédites correspondent parfaitement aux proportions cumulées réelles,plus la courbe de fréquence cumulée se rapproche de la ligne d'égalité parfaite, meilleure est la performance du modèle de fréquence. Si la courbe de fréquence cumulée

se situe en dessous de la ligne d'égalité, cela indique une sous-estimation des fréquences de certaines classes, tandis que si elle se situe au-dessus, cela indique une surestimation. Une courbe qui se rapproche de l'angle supérieur gauche (coin supérieur gauche) indique une meilleure performance du modèle de fréquence, où les fréquences des classes sont correctement prédites avec une plus petite proportion des échantillons.

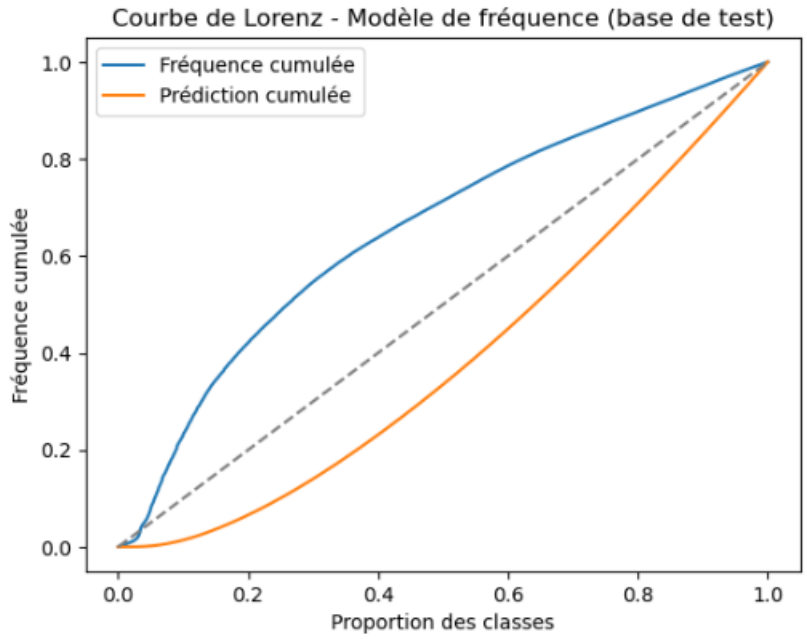


FIGURE 2.17 – Courbes de Lorenz pour le modéle de fréquence appliqué à la base test

la courbe de Lorenz représente le modèle de fréquence (base de test),plus la courbe de fréquence cumulée se rapproche de la ligne d'égalité parfaite au desus et prédiction cumulée au bas de l'égalité parfaite, meilleure est la performance du modèle de fréquence. Si la courbe de fréquence cumulée se situe en dessous de la ligne d'égalité, cela indique une sous-estimation des fréquences de certaines classes, tandis que si elle se situe au-dessus, cela indique une surestimation. Une courbe qui se rapproche de l'angle supérieur gauche (coin supérieur gauche) indique une meilleure performance du modèle de fréquence, où les fréquences des classes sont correctement prédites avec une plus petite proportion des échantillons.

voici l'algorithme et le tableau d'Extrait des observations et de quelques variables du fichier de coût moyen.

```
import pandas as pd

# Chargement des données
production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')

# Sélection d'un échantillon d'observations
sample = production.sample(n=20, random_state=42)  # Changer la valeur de n selon vos besoins

# Sélection des variables spécifiques
selected_variables = ['SEXE_CONDUCTEUR', 'Age_conducteur', 'Prime']
sample_selected = sample[selected_variables]

# Affichage de l'extrait d'observations et de variables
print(sample_selected)
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_7956\2816944020.py:4: DtypeWarning: Columns (8) have mixed types. Specify dtype opti
on on import or set low_memory=False.
  production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')
```

	SEXE CONDUCTEUR	Age conducteur	Prime
450455	M	44	2813.2300
167421	M	60	1614.1700
204374	M	61	6072.2800
517605	M	47	10953.1700
333427	M	30	5085.1504
259386	M	45	2612.9000
236803	M	72	17712.9000
467585	M	48	2004.7400
319706	M	30	12458.0400
509697	M	31	14823.7200
423034	M	42	949.1600
31302	M	57	24261.4300
481372	M	35	5015.3500
393897	M	38	3379.3300
76862	M	58	5058.7300
152828	M	34	12199.1100
468166	M	43	3531.1100
62340	M	57	4496.1300
524367	M	37	2920.6900
141164	M	43	4121.3600

TABLE 2.2 – tableau d'Extrait des observations et de quelques variables du fichier de coût moyen

Nous affichons ci-dessous l'algorithme et leur graphe de correlation entre l'âge du conducteur et la prime obtenus :

```
import pandas as pd
import matplotlib.pyplot as plt

# Chargement des données
production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')

# Sélection d'un échantillon d'observations
sample = production.sample(n=200, random_state=42)  # Changer la valeur de n selon vos besoins

# Sélection des variables spécifiques
selected_variables = ['Age_conducteur', 'Prime']
sample_selected = sample[selected_variables]

# Tracé du scatter plot
plt.scatter(sample_selected['Age_conducteur'], sample_selected['Prime'])
plt.xlabel('Age_conducteur')
plt.ylabel('Prime')
plt.title('correlation entre l\'âge du conducteur et la prime')
plt.show()
```

```
C:\Users\user\AppData\Local\Temp\ipykernel_7956\3208602590.py:5: DtypeWarning: Columns (8) have mixed types. Specify dtype opti
on on import or set low_memory=False.
  production = pd.read_csv('production.csv', delimiter=',', encoding='UTF-8')
```

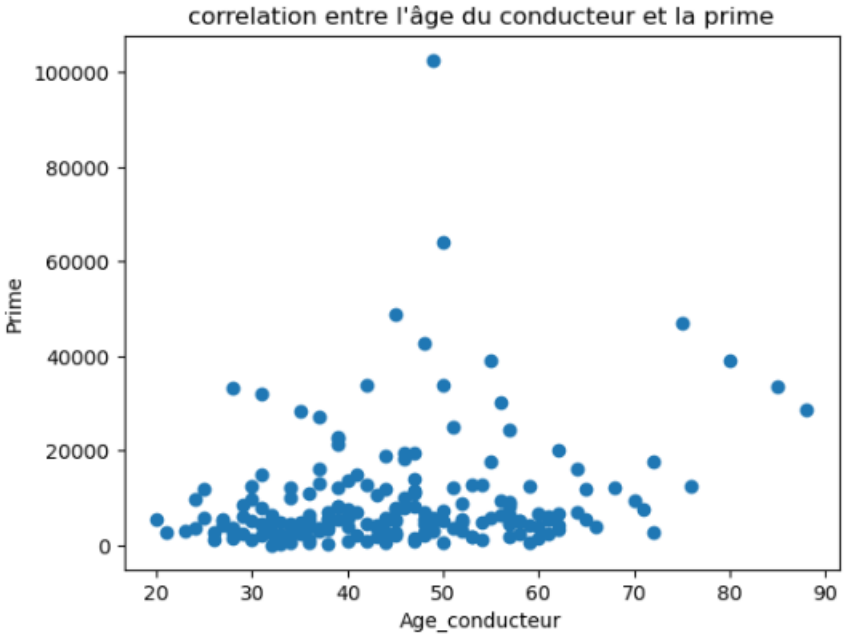


FIGURE 2.18 – correlation entre l'âge du conducteur et la prime

En observant le graphique représentant la corrélation entre l'âge du conducteur et la prime, nous constatons que les points sont uniformément répartis sans schéma distinct de l'âge de 20 ans à 60 ans. Cela suggère qu'il n'y a pas de corrélation évidente entre l'âge du conducteur et la prime dans cette plage d'âge. Cependant, à partir de l'âge de 70 ans jusqu'à 90 ans, les points sont alignés de manière linéaire croissante ou décroissante, ce qui suggère respectivement une corrélation positive ou négative entre les variables.

Les résultats concernant l'extrait d'observation de Sinistre sont affiché ci dessous :

```
import pandas as pd

# Chargement des données
DECLARATION = pd.read_csv('DECLARATION.csv', delimiter=',', encoding='UTF-8')
DECLARATION.groupby('Annee')['ID_S'].nunique()   #NOMBRE DE DECLARATION DANS CHAQUE EXERCICE
DECLARATION.groupby('Annee')['ID'].nunique()   #NOMBRE DE contrat DANS CHAQUE EXERCICE
```

dtype	Annee	ID S	ID
int64	2017	18652	14631
int64	2018	19200	14669
int64	2019	19545	14811
int64	2020	14562	11390
int64	2021	14535	11062

TABLE 2.3 – tableau pour l'extrait d'observation de Sinistre

le graphe pour l'extrait d'observation de Sinistre :

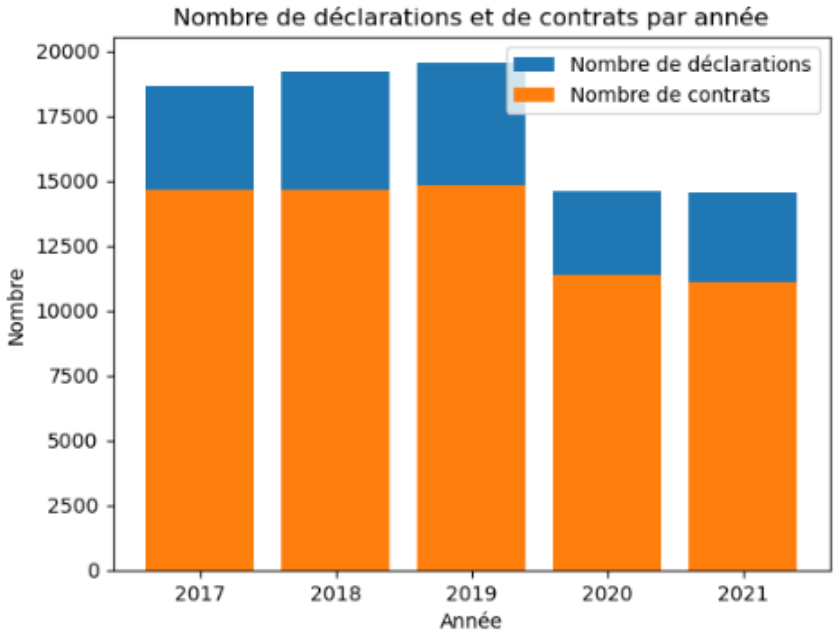


FIGURE 2.19 – Nombre de déclarations et de contrats par année

Le graphique à barres permet de visualiser la variation du nombre de déclarations et de contrats au fil des années. Chaque barre représente une année spécifique, et la hauteur de la barre indique le nombre correspondant de déclarations ou de contrats. on peut tirer des conclusions sur les fluctuations des déclarations et des contrats au fil des années et identifier les périodes où il y a eu une augmentation ou une diminution significative. Cela peut aider à prendre des décisions éclairées sur les politiques ou les actions à entreprendre en fonction des tendances observées.

En examinant les données, nous pouvons observer une tendance générale à la hausse du nombre d'observations de sinistres d'une année à l'autre, de 2017 à 2018. Cela suggère une augmentation des incidents signalés au fil du temps. Cependant, les années 2019 et 2020 se démarquent par des valeurs légèrement inférieures aux années précédentes, ce qui peut être attribué à l'impact de la pandémie de COVID-19.
La pandémie de COVID-19, qui a commencé en 2019 et a eu un impact mondial significatif en 2020, a probablement entraîné une diminution des sinistres signalés. Les restrictions de déplacement, les fermetures d'entreprises et d'autres mesures prises pour limiter la propagation du virus ont pu réduire les risques d'incidents

et, par conséquent, le nombre d'observations de sinistres.
En 2021, nous observons une légère baisse supplémentaire par rapport à l'année précédente, mais cette diminution peut être due à d'autres facteurs indépendants de la pandémie.
voici l'algorithme et le diagramme d'année :

```python
import pandas as pd
import matplotlib.pyplot as plt

# Chargement des données
DECLARATION = pd.read_csv('DECLARATION.csv', delimiter=',', encoding='UTF-8')

# Sélection d'un échantillon d'observations
sample = DECLARATION.sample(n=10, random_state=42)  # Changer la valeur de n selon vos besoins

# Sélection de la variable spécifique
selected_variable = 'Annee'
data = sample[selected_variable]

# Compter les occurrences de chaque valeur
value_counts = data.value_counts()

# Création du diagramme à barres
plt.bar(value_counts.index, value_counts.values)
# Ajout des étiquettes et du titre
plt.xlabel(selected_variable)
plt.ylabel('Fréquence')
plt.title('Diagramme à barres de {}'.format(selected_variable))

# Affichage du diagramme à barres
plt.show()
```

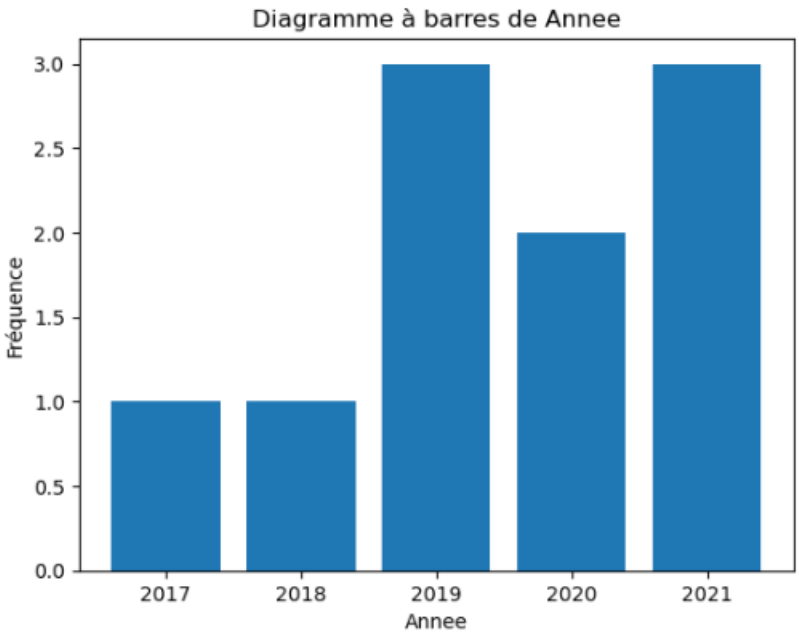


FIGURE 2.20 – Diagramme à barres d'année

ce graphe représente un Diagramme à barres d'annee dépendra des valeurs et des fréquences affichées. Il permet de visualiser la distribution des valeurs de la

variable 'Annee' dans l'échantillon spécifique de données de déclaration. Chaque barre représente une valeur unique et sa hauteur indique la fréquence respective. Cela permet d'identifier les valeurs les plus fréquentes et de détecter des schémas ou des tendances éventuelles dans la répartition des valeurs ,en remarque que l'année 2017-2018 sont de même fréquence et aussi 2019-2021 sont de même fréquence mais 2020 sont fréquence est différente .
La fréquence des années 2017-2018 sont les même, de même que celle des années 2019-2021. Cependant, la fréquence de l'année 2020 est différente.

Arbre de déscision :

Un arbre de décision est un modèle d'apprentissage automatique utilisé pour prendre des décisions ou effectuer des prédictions en utilisant un flux logique basé sur les caractéristiques d'un ensemble de données. Il est souvent utilisé dans des problèmes de classification et de régression.

L'arbre de décision se compose de nœuds qui représentent les caractéristiques ou les attributs des données, de branches qui relient les nœuds et de feuilles qui représentent les résultats ou les prédictions. L'arbre est construit de manière hiérarchique en utilisant des critères de division qui séparent les données en fonction des valeurs des attributs.
L'objectif principal de l'arbre de décision est de créer un modèle prédictif qui peut être utilisé pour prendre des décisions en fonction des caractéristiques des données d'entrée. Il permet de représenter visuellement et de manière intuitive le processus de prise de décision, en utilisant des règles logiques simples pour déterminer le chemin à suivre dans l'arbre en fonction des valeurs des attributs.

L'arbre de décision offre plusieurs avantages, tels que sa simplicité d'interprétation, sa capacité à gérer des données manquantes et à gérer à la fois des variables continues et catégorielles. Cependant, il peut également être sensible aux variations dans les données d'entrée et peut facilement surajuster les données d'entraînement si des précautions ne sont pas prises.

En résumé, un arbre de décision est un modèle d'apprentissage automatique qui utilise une représentation en forme d'arbre pour prendre des décisions ou effectuer des prédictions en fonction des caractéristiques des données d'entrée.

voici ci joint l'arbre de décision d'algorithme CART :

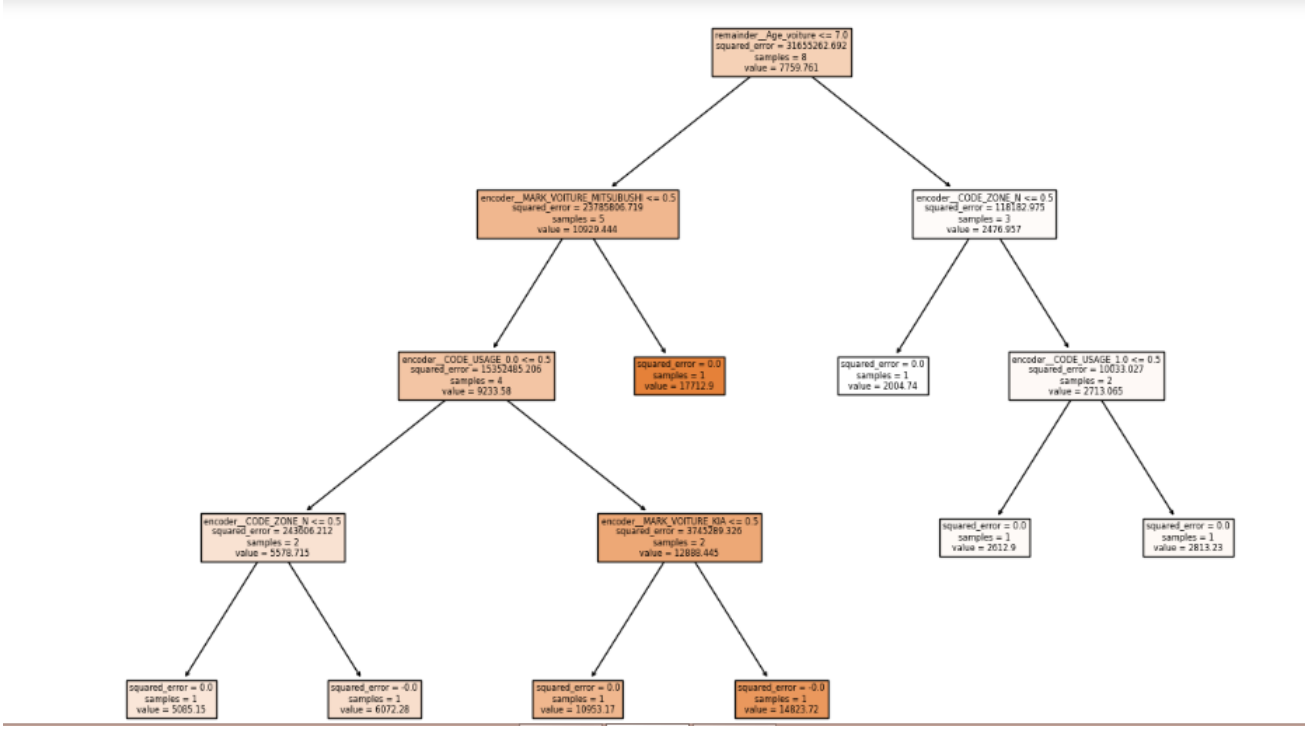


FIGURE 2.21 – L'arbre de décision

Annexe A

La présentation de la SAA :

A.0.1 Introduction :

L'assurance automobile est une forme d'assurance qui offre une protection financière en cas de dommages ou de pertes liés à la conduite d'un véhicule. Elle est obligatoire dans la plupart des pays pour les conducteurs qui utilisent leur véhicule sur la voie publique. L'assurance automobile peut couvrir une variété de risques, tels que les accidents de la route, les dommages causés par des événements naturels ou des actes de vandalisme, le vol de véhicules et la responsabilité civile en cas de dommages corporels ou matériels causés à des tiers.
Les polices d'assurance automobile peuvent être personnalisées pour répondre aux besoins individuels des conducteurs, en fonction de facteurs tels que le type de véhicule, l'âge et l'expérience du conducteur, le lieu de résidence, le niveau de couverture souhaité et le montant de la prime d'assurance. Les types de couverture disponibles peuvent varier selon les pays et les réglementations locales.
En résumé, l'assurance automobile est une forme d'assurance qui offre une protection financière pour les conducteurs en cas de dommages ou de pertes liés à la conduite d'un véhicule, et elle est généralement obligatoire dans la plupart des pays.

La SAA a augmenté son capital social à 35 milliards DA. Une décision importante qui vient couronner plusieurs décennies de succès de l'entreprise, qui célèbre cette année ses 60 ans d'existence. C'est là une démarche stratégique témoignant de la solidité financière de la compagnie et lui permettant de réaffirmer sa position de leader sur le marché.

A.0.2 Définition d'assurance :

L'assurance automobile est une forme d'assurance qui fournit une protection financière en cas d'accident de voiture, de vol, d'incendie ou d'autres types de dommages liés à votre véhicule. Les propriétaires de voitures paient généralement une prime périodique à une compagnie d'assurance automobile, et en retour, l'assureur couvrira les coûts associés aux dommages subis par la voiture. Les types de couverture d'assurance automobile comprennent la responsabilité civile, la collision, la protection contre les conducteurs non assurés ou sous-assurés, la protection contre les dommages causés par la nature (par exemple, les tempêtes), et d'autres options de couverture supplémentaires. Il est généralement obligatoire d'avoir une assurance automobile pour conduire sur la route dans la plupart des pays.

A.0.3 L'organigramme de la société :

Concernant la structure organisationnelle d'une compagnie d'assurance automobile, elle peut varier en fonction de la taille de la compagnie et de son modèle d'entreprise. En général, cependant, les compagnies d'assurance automobile sont dirigées par un conseil d'administration, qui est chargé de superviser la gestion de la société et de prendre des décisions importantes. Le conseil d'administration nomme également un directeur général, qui est responsable de la gestion quotidienne de l'entreprise.
En outre, les compagnies d'assurance automobile peuvent être divisées en plusieurs divisions, chacune ayant ses propres responsabilités et objectifs. Par exemple, une division peut être chargée de la souscription des polices d'assurance, une autre de la gestion des sinistres, et une autre encore de la commercialisation des produits d'assurance. Chaque division est généralement dirigée par un responsable, qui relève du directeur général.
La société nationale d'assurance SAA est composé de :

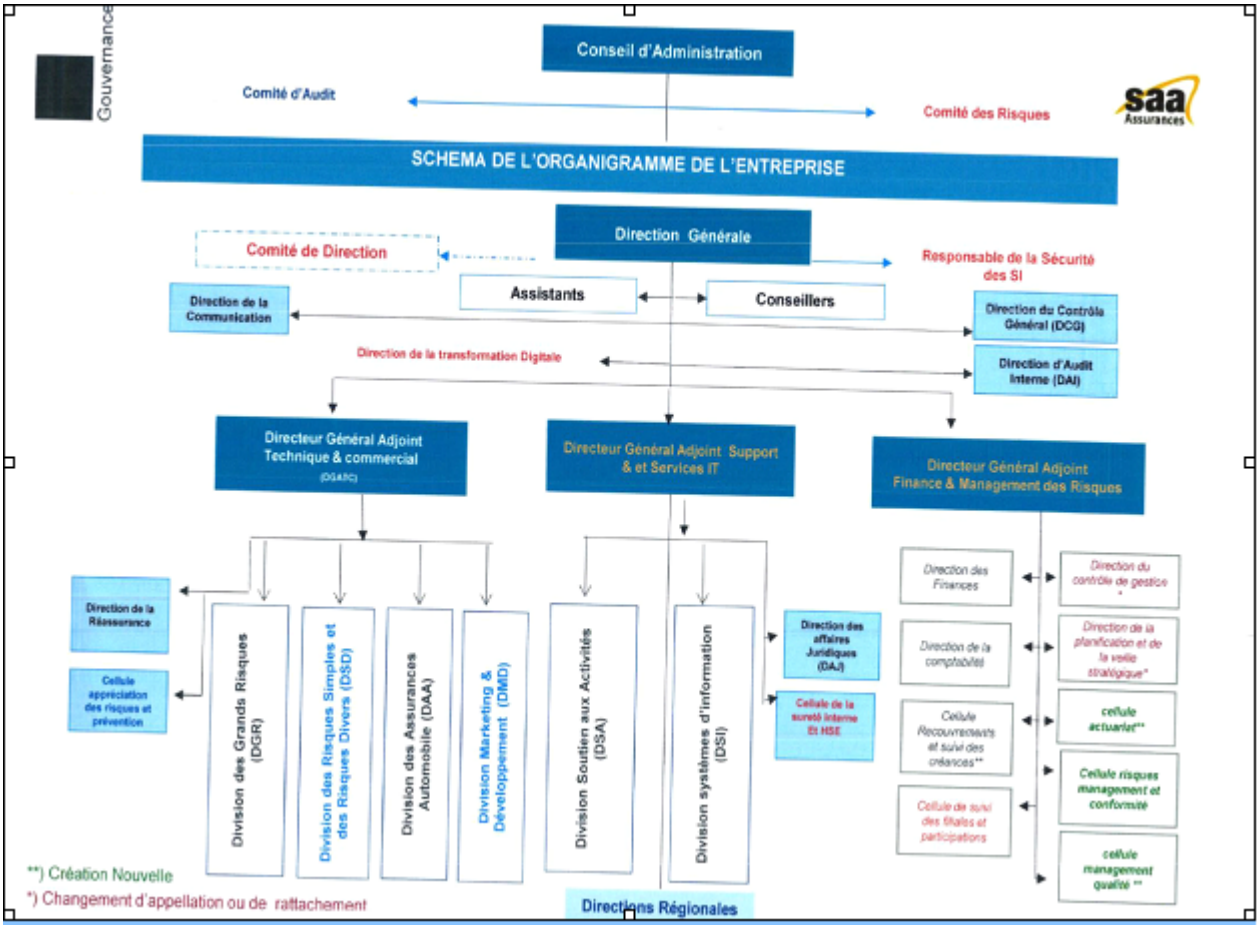


FIGURE A.1 – Organigramme de la SAA

Conseil d'administration :

Le conseil d'administration est responsable de superviser la gestion de la compagnie et de prendre des décisions importantes. Le conseil est composé de membres externes à la compagnie et peut inclure des investisseurs, des représentants du secteur financier, des experts juridiques et des personnalités du monde des affaires.

Comité d'audit :

sert à surveiller la gestion financière de l'entreprise. Généralement composé de membres indépendants du conseil d'administration et de la direction de l'entreprise, qui ont une expérience et une expertise en comptabilité, finance et audit. Son rôle est d'assurer la conformité de l'entreprise aux normes financières et légales, à la gestion des risques financiers et à la gestion de la qualité de l'information financière. Il permet d'examiner les états financiers de l'entreprise, les rapports d'audit, les contrôles internes et les politiques comptables, et fournit des recommandations pour améliorer les pratiques de gestion financière de l'entreprise.

Comité de risque :

sert à surveiller et de gérer les risques de l'entreprise. Il est généralement composé de membres du conseil d'administration et de la direction de l'entreprise, qui ont une expérience et une expertise en gestion de risques. Le comité de risque travaille à identifier les risques potentiels pour l'entreprise, à évaluer l'impact de ces risques sur l'entreprise et à développer des plans d'atténuation des risques. Il permet d'examiner les risques liés à l'ensemble des activités de l'entreprise, notamment les risques financiers, opérationnels, de conformité et de cyber sécurité. Il fournit des recommandations pour améliorer la gestion des risques de l'entreprise et pour s'assurer que les risques sont gérés de manière efficace et efficiente.

Direction générale :

Le directeur général est nommé par le conseil d'administration pour gérer la compagnie au quotidien. Il est responsable de la mise en œuvre des politiques de la compagnie et de la réalisation des objectifs commerciaux.

Comité de direction :

Le comité de direction est composé des principaux dirigeants de la compagnie, y compris le PDG, les DGA et les directeurs de division. Il est responsable de la prise de décision stratégique et de la gestion globale de la compagnie.

Responsable de la sécurité des SI :

Le responsable de la sécurité des SI est responsable de la sécurité des systèmes informatiques de la compagnie. Il travaille en étroite collaboration avec la division

des systèmes d'information pour s'assurer que les systèmes sont sécurisés contre les cyberattaques.
Les assistants et conseillers sont des membres de l'équipe de direction qui travaillent en étroite collaboration avec les dirigeants de l'entreprise pour les aider dans leurs tâches quotidiennes et les conseiller dans leur prise de décisions.

Les assistants de direction :

sont généralement chargés de tâches administratives et de soutien, telles que la gestion des agendas, la planification des réunions, la rédaction de rapports et la gestion des communications. Ils peuvent également être responsables de la coordination des activités de plusieurs divisions ou départements, ou de la gestion de projets spécifiques.

Les conseillers :

quant à eux, sont des experts dans leur domaine qui fournissent des conseils et des recommandations aux dirigeants de l'entreprise pour les aider à prendre des décisions éclairées. Les conseillers peuvent être des consultants externes ou des employés de l'entreprise, et ils peuvent avoir une expertise dans divers domaines, tels que la finance, la stratégie d'entreprise, les ressources humaines, la technologie ou le marketing. Leur rôle est d'analyser les données et les tendances, de fournir des conseils basés sur leur expertise, et de collaborer avec les dirigeants de l'entreprise pour élaborer des stratégies et des plans d'action.

Direction de la communication :

La direction de la communication est responsable de la gestion de la communication externe de la compagnie, y compris la publicité, les relations publiques et les réseaux sociaux.
La Direction des Contrôles Généraux et la Direction d'Audit Interne sont deux fonctions distinctes mais complémentaires dans la gouvernance d'entreprise.

La Direction des Contrôles Généraux (DCG) :

a pour rôle de veiller à ce que les processus et les opérations de l'entreprise soient conformes aux règles, réglementations et politiques internes de l'entreprise.

Elle évalue l'efficacité et l'efficience des systèmes de contrôle interne et des processus opérationnels, afin d'identifier les risques potentiels et de proposer des recommandations pour améliorer les processus et réduire les risques. La DCG est généralement responsable de la mise en œuvre et de la gestion des politiques et procédures de contrôle interne de l'entreprise, ainsi que de la surveillance et de l'évaluation de leur efficacité.

La Direction d'Audit Interne (DAI) :

est chargée d'évaluer la qualité et l'efficacité des contrôles internes de l'entreprise, ainsi que de la gestion des risques et de la conformité aux lois et réglementations. Elle examine les processus, les opérations et les systèmes de l'entreprise pour évaluer l'efficacité et l'efficience de leur fonctionnement et pour identifier les risques potentiels. La DAI travaille également en collaboration avec la Direction des Contrôles Généraux pour élaborer des plans d'audit et pour effectuer des audits réguliers afin d'assurer la conformité de l'entreprise aux normes internes et externes.

Directeur général adjoint (DGA) en charge de la communication technique et commerciale :

Le DGA en charge de la communication technique et commerciale est responsable de la stratégie de communication de la compagnie. Il s'occupe notamment de la communication avec les clients, la presse et les partenaires commerciaux. Il travaille également en étroite collaboration avec les autres divisions pour développer des produits et des services qui répondent aux besoins des clients .qui contient :

*** Division des grands risques :** Cette division est responsable de la gestion des risques élevés et des sinistres majeurs. Elle gère les polices d'assurance des clients qui présentent un risque plus élevé que la moyenne, tels que les conducteurs avec un historique de sinistres, les conducteurs d'âge avancé ou les conducteurs qui possèdent des voitures haut de gamme. Cette division a des experts en analyse de risques qui travaillent en étroite collaboration avec les autres divisions pour évaluer les risques et fixer des primes d'assurance appropriées.

*** Division des simples risques :** Cette division gère les polices d'assurance pour les clients présentant un risque moins élevé que la moyenne, tels que les conducteurs ayant un bon dossier de conduite et une voiture plus ancienne ou

moins chère. Cette division est chargée de la souscription des polices et de la fixation des tarifs d'assurance.

*** Division des assurances automobiles :** Cette division est responsable de la conception, de la tarification et de la gestion des polices d'assurance automobile. Elle gère également les réclamations liées à l'assurance automobile et travaille en étroite collaboration avec la division des grands risques et des simples risques pour déterminer les tarifs approprié.

*** Division marketing :** Cette division est chargée de la promotion des produits d'assurance de la compagnie. Elle est responsable de la conception des produits, de la tarification, de la promotion et de la distribution des produits d'assurance. Le responsable de cette division s'assure que la compagnie offre des produits d'assurance adaptés aux besoins du marché et qu'elle atteint ses objectifs de vent.

Directeur général adjoint (DGA) support et service IT :

Le DGA en charge du support et du service IT est responsable de la gestion des systèmes informatiques et de l'infrastructure technologique de la compagnie. Il veille à ce que les systèmes d'information soient disponibles et fonctionnent correctement pour permettre aux autres divisions de remplir leur mission. Il travaille également en étroite collaboration avec la division des systèmes d'information pour développer de nouveaux systèmes et améliorer l'efficacité des systèmes existants.qui contient :

*** Direction des affaires juridiques :** La direction des affaires juridiques est responsable de la gestion des questions juridiques et réglementaires de la compagnie. Elle conseille la direction sur les questions juridiques et travaille en étroite collaboration avec les autres divisions pour s'assurer que la compagnie est en conformité avec les lois et règlements applicables.

*** Division de soutien aux activités :** Cette division fournit des services de soutien administratif aux autres divisions, tels que la gestion des ressources humaines, la comptabilité, les achats et la gestion des installations. Elle s'assure que les processus internes sont efficaces et que les employés disposent des ressources nécessaires pour remplir leur mission.

*** Division des systèmes d'information :** Cette division est chargée de la ges-

tion des systèmes informatiques et de l'infrastructure technologique de la compagnie. Elle est responsable du développement, de la mise en œuvre et de la maintenance des systèmes informatiques utilisés par les autres divisions, tels que les systèmes de gestion des sinistres, de souscription et de comptabilité. Elle est également responsable de la sécurité des systèmes d'information et de la prévention des cyberattaques.

*** Direction régionale :** Les directions régionales sont responsables de la gestion des opérations de la compagnie dans une région géographique donnée. Elles travaillent en étroite collaboration avec les autres divisions pour s'assurer que les activités de la compagnie sont adaptées aux besoins locaux.

Directeur général adjoint finance et management des risques :

son rôle dans une compagnie d'assurance automobile est essentiellement de superviser et de diriger les activités liées à la gestion financière et à la gestion des risques de la société. Les principales responsabilités de ce poste peuvent inclure :

*** Planification financière :** Le directeur général adjoint finance et management des risques doit travailler en étroite collaboration avec d'autres membres de la direction pour établir des objectifs financiers et stratégiques pour la société. Il est chargé de la planification budgétaire, de la prévision financière et de l'évaluation des risques financiers.

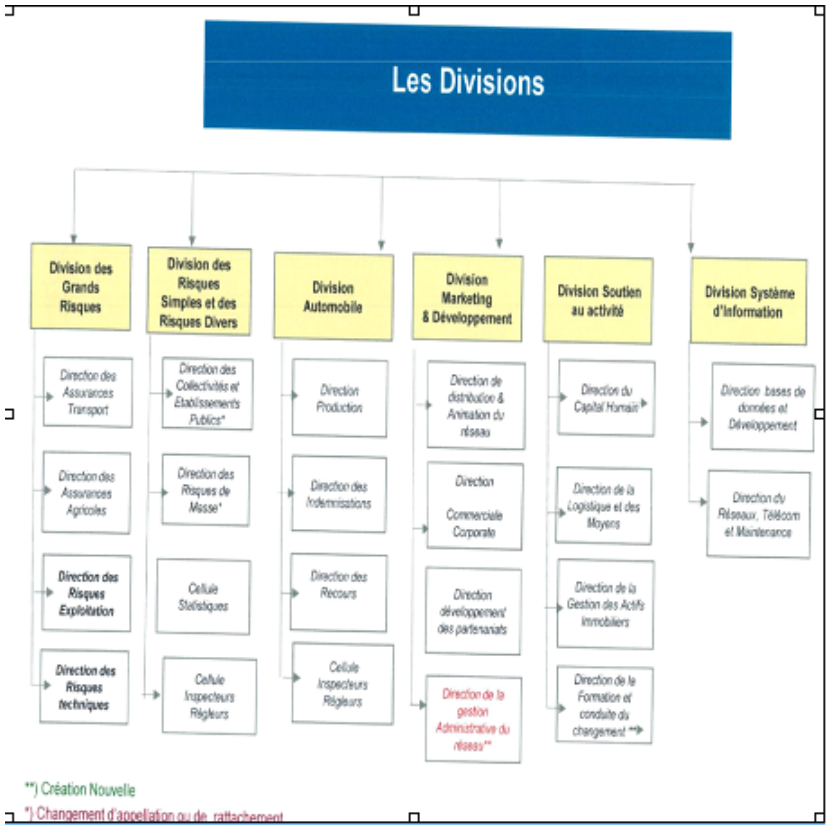
*** Gestion des risques :** Le directeur général adjoint finance et management des risques doit évaluer les risques encourus par la société et élaborer des stratégies pour les gérer de manière efficace. Cela peut inclure l'analyse des risques liés aux investissements, la mise en place de politiques de couverture de risques et la surveillance de la conformité aux réglementations financières.

***Le directeur général adjoint finance et management des risques :** est responsable de la production de rapports financiers précis et opportuns, à la fois pour la direction et pour les parties prenantes externes. Cela peut inclure des rapports annuels, trimestriels et mensuels, ainsi que des rapports sur les performances financières.
*** Leadership d'équipe :** Le directeur général adjoint finance et management des risques doit diriger une équipe de professionnels de la finance et de la gestion des risques, en les guidant dans l'élaboration et la mise en œuvre de stratégies

efficaces. Il est également responsable de la formation et du développement professionnel de l'équipe.

*** Leadership d'équipe :** Le directeur général adjoint finance et management des risques doit diriger une équipe de professionnels de la finance et de la gestion des risques, en les guidant dans l'élaboration et la mise en œuvre de stratégies efficaces. Il est également responsable de la formation et du développement professionnel de l'équipe.

Ces divisions peuvent être subdivisées en unités plus petites selon les besoins de la compagnie. Chaque division ou unité peut avoir un responsable qui rapporte au directeur général de la compagnie.



Conclusion générale

Dans le cadre de cette étude sur la tarification en assurance automobile, nous avons exploré deux approches distinctes : les modèles linéaires généralisés (GLM) et l'algorithme CART pour la classification des résidus. Notre objectif principal était d'évaluer l'efficacité de ces méthodes dans le contexte de l'assurance automobile de la SAA.

Lors de notre étude de pratique dans le contexte de l'assurance automobile de la SAA, nous avons appliqué ces deux algorithmes, à savoir le GLM et le CART. Nous avons pu constater que ces méthodes amélioraient de manière significative la précision des estimations de primes d'assurance. De plus, elles ont permis de détecter des polices d'assurance présentant des caractéristiques particulières qui auraient pu passer inaperçues avec des méthodes traditionnelles.

cette étude a apporté des contributions significatives à la tarification en assurance automobile. Les modèles linéaires généralisés, en utilisant l'algorithme GLM, ainsi que l'algorithme CART pour la classification des résidus, se sont avérés efficaces pour modéliser les relations entre les variables et les primes d'assurance. Ces approches ont permis une meilleure compréhension des facteurs influençant la tarification et ont conduit à des estimations plus précises des primes d'assurance automobile.

Il convient de noter que cette étude comporte certaines limites. Par exemple, nous avons utilisé des données spécifiques de l'assurance automobile de la SAA, et les résultats peuvent varier dans d'autres contextes ou pour d'autres compagnies d'assurance. De plus, il serait intéressant d'explorer d'autres variables ou d'autres

algorithmes pour continuer à améliorer les modèles de tarification.

cette recherche a fourni des perspectives intéressantes pour la tarification en assurance automobile en utilisant des méthodes statistiques avancées. Les résultats obtenus peuvent être utilisés pour améliorer les processus de tarification dans l'industrie de l'assurance automobile, en fournissant des estimations plus précises des primes et en permettant une meilleure gestion des risques. Ces résultats ouvrent également la voie à de futures recherches dans le domaine de la tarification en assurance automobile.

D'après les résultats du test BIC (Bayesian Information Criterion), la meilleure loi est la loi Gamma, car elle présente un BIC plus faible par rapport à la loi gaussienne. Le test BIC est utilisé pour évaluer la qualité de l'ajustement d'un modèle statistique en prenant en compte à la fois l'ajustement aux données et la complexité du modèle. Dans ce cas, un BIC plus faible indique une meilleure adéquation du modèle aux données.
En utilisant un jeu de données comprenant des variables telles que l'âge du conducteur, l'année, la puissance de la voiture et l'âge de la voiture, nous avons comparé les performances des méthodes GLM (Generalized Linear Model) et CART (Classification and Regression Trees) pour prédire la prime d'assurance automobile.Après avoir entraîné les modèles sur un ensemble d'entraînement et effectué des prédictions sur un ensemble de test, nous avons évalué leurs performances à l'aide du RMSE (Root Mean Squared Error), qui mesure l'écart moyen entre les valeurs prédites et les valeurs réelles.
Les résultats obtenus ont montré que le modèle GLM a un RMSE de 12903.448 , tandis que le modèle CART a un RMSE de 10791.620.
En comparant ces résultats, nous constatons que le modèle CART présente un RMSE inférieur, ce qui indique une meilleure performance en termes de précision de prédiction de la prime d'assurance automobile.

RMSE

Le RMSE (Root Mean Squared Error) est une mesure d'erreur utilisée pour évaluer les performances d'un modèle de régression. Il mesure l'écart moyen entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible, en prenant en compte la racine carrée de l'erreur quadratique moyenne.
voici l'algorithme et l'éxecusion de teste de deux méthode GLM et CART :

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor

# Charger les données dans un DataFrame
data = pd.DataFrame(BASE)

# Colonnes à convertir en numérique
colonnes_numeric = ['Unnamed: 0', 'Annee', 'Age_conducteur', 'PUISS_VOITURE', 'Age_voit

# Convertir les colonnes en numérique
data[colonnes_numeric] = data[colonnes_numeric].apply(pd.to_numeric, errors='coerce')

# Remplacer les NaN par des valeurs numériques (0)
data = data.fillna(0)

# Diviser les données en ensembles d'entraînement et de test
X = data[['Age_conducteur', 'Annee', 'PUISS_VOITURE', 'Age_voiture']]
y = data['Prime']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=4

# Appliquer l'algorithme GLM
formula = "Prime ~  Age_conducteur + Annee + PUISS_VOITURE + Age_voiture"
model = sm.GLM.from_formula(formula, data=data, family=sm.families.Binomial())
results = model.fit()

# Prédictions du modèle GLM
glm_predictions = results.predict(data)

# Évaluation des performances du modèle GLM
glm_rmse = mean_squared_error(data['Prime'], glm_predictions, squared=False)

# Créer et entraîner le modèle CART
cart_model = DecisionTreeRegressor()
cart_model.fit(X_train, y_train)
```

```
# Prédictions du modèle CART sur l'ensemble de test
cart_predictions = cart_model.predict(X_test)

# Évaluation des performances du modèle CART
cart_rmse = mean_squared_error(y_test, cart_predictions, squared=False)

# Comparaison des performances
print("GLM - RMSE:", glm_rmse)
print("CART - RMSE:", cart_rmse)
```

```
GLM - RMSE: 12903.44803437077
CART - RMSE: 10791.620181905135
```

Bibliographie

[1] André Bayala. Assurance et développement durable. Revue d'économie financière, pages 317–327, 2005.

[2] Bouaziz Cheikh. L'histoire de l'assurance en algérie. Assurances et gestion des risques, vol81 (3-4) Octobre-Décembre, 2013.

[3] Brahim GUENANE. Mesurer la satisfaction clients par les méthodes précises etude de cas de la compagnie algérienne d'assurance et de réassurance (caar). MO'assira Economic Research, 4(1) :119–134, 2021.

[4] Elkhansa M'ahammedi and Samira Tellache. Modélisation du coût des sinistres en assurance automobile. PhD thesis, 2020.

[5] Francois Couibault. Constant eliashberg, and michel latrasse, 2003. Les grands principes de rassurance.

[6] François Couilbault, Stéphanie Couilbault-Di Tommaso, and Nadia Hadj-Chaib Candeille. Les grands principes de l'assurance. L'Argus de l'assurance éditions, 2023.

[7] Guillaume Gonnet. Etude de la tarification et de la segmentation en assurance automobile. Université de Lyon, Université Claude Bernard–Lyon, 1, 2010.

[8] Jacques Charbonnier. Dictionnaire de la gestion des risques et des assurances. La Maison du dictionnaire, 2004.

[9] Jean-Pierre Lemaire. Tarification en assurance : évaluation des risques et fixation des primes. Revue économique et financière, 45(2) :123–138, 2020.

[10] Lamia Bennamane. L'apport de l'assurance des risques industriels dans la performance financière d'une compagnie d'assurance : Cas de la SAA. PhD thesis, Université Mouloud Mammeri, 2023

[11] LE Julien. Titre : Tarification d'un produit assurance automobile au tiers mobilisant des données de marché.

[12] Magali Ruimy. Elaboration d'un véhiculier en assurance automobile. Mémoire ISFA, 2017.

[13] Marie-Christine Blais. La tarification en assurance automobile : facteurs et méthodes. Revue de l'assurance automobile, 36(2) :87–102, 2018.

[14] Massil Ammarkhoudja and Ania Amoura. L'Analyse Financière dans une société d'assurance publique (Cas société algérienne d'assurance SAA Bouzguene). PhD thesis, Université Mouloud Mammeri, 2021.

[15] Michel Denuit. Le rôle de la tarification en assurance. Risques, 102(3) :79–98, 2015.

[16] Paul Durand. La relation entre automobile et tarification assurance : facteurs et impact sur les primes d'assurance. Revue de l'assurance, 42(4) :217–232, 2021.

[17] Pierre-Yves Geoffard. Les bases de la tarification. Revue économique, 60(3) :553–578, 2009.

[18] Rémi Bellina. Méthodes d'apprentissage appliquées à la tarification non-vie. Mémoire ISFA, 2014.

[19] Sophie Martin. Les types de tarification en assurance automobile : analyse et comparaison. Revue de l'assurance automobile, 40(3) :127–142, 2019.