Diplôme d'Université
Big data, Data science
Analyse des risques sous Python

Université de Montpellier

Analyse de Données:
Optimisation des Stratégies d'Assurance par le Traitement Avancé de Données

Présenté Par :

Samadou KODON
Abdoulaye SOW
Daouda LY

Sous la direction du professeur:

Gilles Michel
Samuel.STOCKSIEKER

Année Scolaire : 2023 /2024

Table des matières

Introduction

Dans le secteur dynamique et en constante évolution de l'assurance, les insurtechs comme la nôtre jouent un rôle crucial en innovant à travers l'utilisation des données pour devancer la concurrence et répondre efficacement aux exigences réglementaires changeantes. Dans ce contexte, l'analyse de données avancée ne se présente plus comme un luxe mais comme une impérieuse nécessité stratégique.

Maîtriser ces techniques est essentiel pour convertir les volumes considérables de données brutes en connaissances précieux qui orientent les décisions critiques et maximisent les ressources.

Ce rapport documente notre exploration stratégique des données, visant à mieux comprendre et anticiper les comportements des assurés et à affiner les stratégies de tarification pour rester à la pointe des attentes fluctuantes des clients et des tendances du marché.

Utilisant Python comme notre principal outil d'analyse, nous avons exploité des ensembles de données pour révéler des schémas significatifs et découvrir des opportunités inédites ainsi que des risques non observés initialement.

Ce projet vise à utiliser ces connaissances pour répondre de manière proactive aux besoins de notre équipe et de notre hiérarchie, en améliorant la prise de décisions et en optimisant les performances opérationnelles.

Nous débuterons par définir les objectifs spécifiques de cette analyse, suivis de notre méthodologie détaillée. Nous mènerons ensuite une analyse exploratoire pour déterminer comment les variables explicatives influencent le montant des sinistres. Des modèles économétriques et de machine learning seront déployés et les résultats seront soigneusement examinés. Enfin, nous discuterons des limites de notre approche et proposerons des orientations pour des recherches futures.

Objectif du Projet

L'objectif principal de ce projet est d'exploiter les capacités avancées de l'analyse de données pour améliorer la compréhension des comportements des assurés et optimiser les stratégies de tarification au sein de notre insurtech.

Nous visons à atteindre cet objectif à travers plusieurs axes spécifiques:

1. Examiner les Données: Nous commencerons par analyser la structure de nos bases de données d'assurance, identifier les informations manquantes et rectifier les erreurs ou les données qui paraissent anormales.

2. Corriger les Erreurs : Après avoir identifié les anomalies, nous les corrigerons pour assurer la fiabilité de nos analyses. Chaque correction sera documentée pour garantir que notre processus est transparent.

3. Explorer les Relations entre Variables: Nous utiliserons des analyses statistiques pour examiner comment différentes variables sont liées entre elles et leur impact sur des éléments importants comme le montant des sinistres.

4. Utiliser des modèles pour prédire les comportements : Nous développerons des modèles statistiques et de machine learning pour prédire les comportements futurs des assurés et identifier les risques potentiels.

5. Interpréter et Utiliser les Résultats : Nous expliquerons ce que nos analyses révèlent et utiliserons ces informations pour formuler des recommandations pratiques pour améliorer notre façon de fixer les prix et de gérer les risques.

2. Examen des Données

2.1. Description des Données Utilisées

Notre étude utilise une base de données d'assurance qui renferme des informations détaillées sur les assurés et leurs sinistres. Cette base comprend des données personnelles, des informations sur les comportements de conduite, ainsi que des détails relatifs aux réclamations effectuées. Ces données fournissent une vue complète et essentielle pour évaluer les risques associés à chaque assuré. Nous avons fourni dans la partie annexe, une descriptions de nos colonnes (Tableau 1)

2.2. Nettoyage Préliminaire et Corrections Effectuées

Le nettoyage initial des données a été réalisé dans Excel, où nous avons entrepris plusieurs étapes clés pour uniformiser et préparer les données à des analyses plus approfondies. Nous avons modifié les virgules en points dans les champs numériques pour uniformiser le format à travers toute la base de données. De plus, des corrections ont été apportées dans la colonne ***NB_Claim***, où des entrées telles que 'NB_CLAIM :1' ont été remplacées par '1', 'NB_CLAIM :2' par '2', et 'NB_CLAIM :3' par '3' pour standardiser ces données.

2.3. Traitements et Corrections avec Python

Après l'importation des données dans Python, des contrôles et corrections supplémentaires ont été nécessaires pour assurer leur qualité et leur cohérence. Le préfixe ***'cnt'*** a été supprimé des identifiants dans la colonne ***Id_Pol*** pour standardiser les numéros de polices. Les valeurs négatives trouvées dans les colonnes ***Pct.drive.sun*** et ***Car_age*** ont été éliminées, car elles ne sont pas logiques dans le contexte des données. Les dossiers des assurés de moins de **18 ans** ont été exclus pour respecter les réglementations légales. Les données manquantes dans les colonnes ***Région*** et ***Marital*** ont été imputées par le mode, et les incohérences dans la colonne ***Car_use*** ont été rectifiées. De plus, nous avons identifié et retiré les doublons parfaits dans notre clé primaire pour éviter toute redondance et garantir l'unicité des données dans notre analyse.

Enfin, nous avons opté pour une jointure interne plutôt qu'une jointure externe pour garantir la qualité et la fiabilité de nos données analysées.Car l'utilisation d'une jointure externe dans notre cas conduirait à un nombre excessivement élevé de valeurs manquantes, ce qui pourrait sérieusement biaiser ou invalider les résultats de notre analyse. En privilégiant une jointure interne, nous sacrifions certes une partie de la quantité des données disponibles, mais nous assurons que toutes les données incluses dans l'analyse sont complètes et fiables. Cette approche permet de maintenir une haute intégrité des données et d'obtenir des insights plus précis et représentatifs, essentiels pour la prise de décision basée sur cette analyse.

3. Analyse Exploratoire des Données

Dans le cadre de notre projet, l'analyse exploratoire des données a été conduite avec soin pour identifier des tendances, des relations, et des anomalies potentielles dans les données d'assurance. Cette étape a été cruciale pour préparer le terrain à des analyses plus complexes et à la modélisation économétrique.

3.1. Statistiques Descriptives des Variables de Conduite

Notre analyse exploratoire des données a révélé des insights intéressants sur le comportement de conduite des assurés, mesuré à travers diverses variables telles que le pourcentage annuel de conduite (Annual.pct.driven), le total des miles parcourus (Total.miles.driven), et les détails des habitudes de conduite quotidiennes et hebdomadaires. En moyenne, les assurés conduisent 87.33% de l'année, parcourant environ 8,733 miles. La conduite est relativement uniforme tout au long de la semaine, avec un léger pic le vendredi (en moyenne 15.74% de conduite), et diminue pendant le weekend, avec une moyenne de 11.12% le dimanche.

La variabilité des comportements de conduite est également reflétée dans les mesures de l'accélération et du freinage, où les assurés ont en moyenne 50.95 événements d'accélération forte et 99.09 événements de freinage fort par 100 miles. Cela indique une tendance à une conduite potentiellement agressive ou réactive. En outre, les intensités des virages à gauche et à droite montrent une activité modérée avec des moyennes significatives, suggérant des habitudes de conduite actives dans les environnements urbains ou complexes (confer tableau 2 en Annexe)

Ces données démontrent non seulement les patterns de conduite parmi les assurés mais fournissent également une base pour comprendre les risques potentiels et pour ajuster les politiques d'assurance en conséquence. La compréhension de ces comportements est cruciale pour optimiser les stratégies de tarification et de gestion des risques, en ciblant des interventions spécifiques selon les habitudes de conduite observées.

3.2. Analyse des Variables Catégorielles

Nous avons utilisé un test du chi-carré pour explorer les liens entre le statut marital et la fréquence des sinistres. Notre statistique de test de 3.6817 et une valeur-p de 0.1587 indiquent qu'il n'y a pas de preuve suffisante d'une association significative entre ces variables. Ainsi, le statut marital ne

semble pas influencer directement la fréquence des sinistres, suggérant que d'autres facteurs pourraient jouer un rôle plus déterminant.

3.3. Analyse de l'Association entre l'Utilisation des Véhicules et la Fréquence des Sinistres

Notre étude a révélé des résultats contrastés lors de l'exploration des effets de l'utilisation des véhicules (Car_use) et de la région (Region) sur les sinistres. Pour Car_use, le test a montré une association forte (statistique de test de 24.3373 et valeur-p de 0.0020), indiquant que la manière dont les véhicules sont utilisés affecte significativement la fréquence des sinistres. Cette découverte est essentielle pour affiner les stratégies de gestion des risques et de tarification. En revanche, pour Region, la statistique de test de 3.5178 avec une valeur-p de 0.1722 n'a pas fourni de preuves suffisantes pour soutenir une association significative, ce qui implique que les variations régionales des sinistres pourraient être dues à d'autres facteurs.

Ces analyses révèlent que l'utilisation du véhicule est un prédicteur clé de la fréquence des sinistres, tandis que la région ne montre pas une corrélation directe. Cette information est cruciale pour cibler les interventions et pour comprendre les dynamiques qui affectent les sinistres selon différentes conditions d'utilisation des véhicules et zones géographiques.

3.4. Analyse des Corrélations entre les Variables de Conduite et le Montant des Sinistres

Notre étude statistique a révélé des relations significatives entre plusieurs variables de conduite et le montant des sinistres, ce qui offre des insights pertinents pour la gestion des risques et l'ajustement des politiques d'assurance. Notamment, nous avons observé que certains comportements de conduite sont fortement corrélés entre eux, ce qui peut influencer le montant des sinistres.

La corrélation de Kendall entre les variables Annual.pct.driven et Total.miles.driven est de 0.4238, indiquant une association modérée. Cela suggère que les véhicules utilisés plus fréquemment tout au long de l'année ont tendance à accumuler plus de miles, ce qui pourrait augmenter le risque de sinistres et, par conséquent, influencer le montant des sinistres associés.

En outre, les variables Left.turn.intensity08 et Left.turn.intensity09 ont montré une très forte corrélation de 0.9347, tandis que Right.turn.intensity10 et Right.turn.intensity11 ont également présenté une corrélation similaire de 0.9285. Ces fortes corrélations indiquent que les mesures d'intensité de virage sont cohérentes et pourraient être utilisées conjointement pour évaluer plus précisément le comportement de conduite agressif ou risqué, qui est souvent un prédicteur de sinistres plus élevés.

Surprenamment, la corrélation entre le montant des sinistres (AMT_Claim) et les autres variables de conduite comme Annual.pct.driven est relativement faible (-0.0024), ce qui suggère que la fréquence d'utilisation annuelle du véhicule ne prédit pas directement les coûts des sinistres. Cela pourrait indiquer que d'autres facteurs, tels que les conditions spécifiques de conduite ou les caractéristiques environnementales, jouent un rôle plus significatif dans la détermination des montants des sinistres.

Cette analyse des corrélations nous permet de mieux comprendre les liens entre les habitudes de conduite et les sinistres. Elle souligne l'importance d'une évaluation plus nuancée des risques, qui devrait prendre en compte non seulement la fréquence et la distance parcourue mais aussi le style de conduite, comme indiqué par les intensités de virage. En intégrant ces variables dans nos modèles de tarification et de gestion des risques, nous pourrions améliorer significativement la précision de nos prédictions de sinistres et optimiser les stratégies d'assurance pour mieux répondre aux besoins spécifiques de nos assurés.

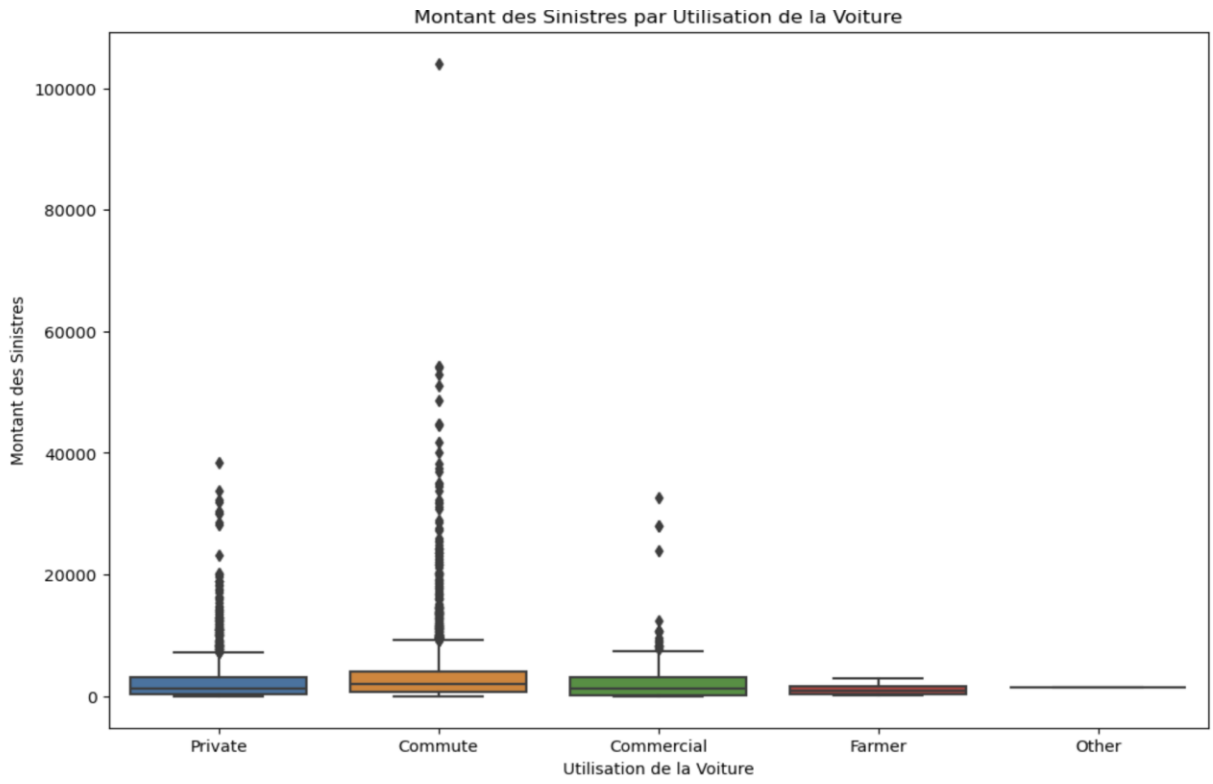
4. Visualisations des données

4.1. Montant des Sinistres par Utilisation de la Voiture

Le graphique ci-dessous illustre clairement que la fréquence et la gravité des sinistres varient considérablement selon l'usage de la voiture. Les véhicules utilisés à des fins communes ou commerciales semblent afficher une variabilité significative dans les montants des sinistres, indiquant potentiellement des risques plus élevés ou des utilisations plus intensives. En revanche, les véhicules utilisés à des fins privées montrent les montants de sinistres les plus faibles et les moins variables, ce qui peut refléter une conduite plus prudente ou moins fréquente.

Cette analyse suggère que l'utilisation du véhicule devrait être un facteur clé dans la modélisation des primes d'assurance.

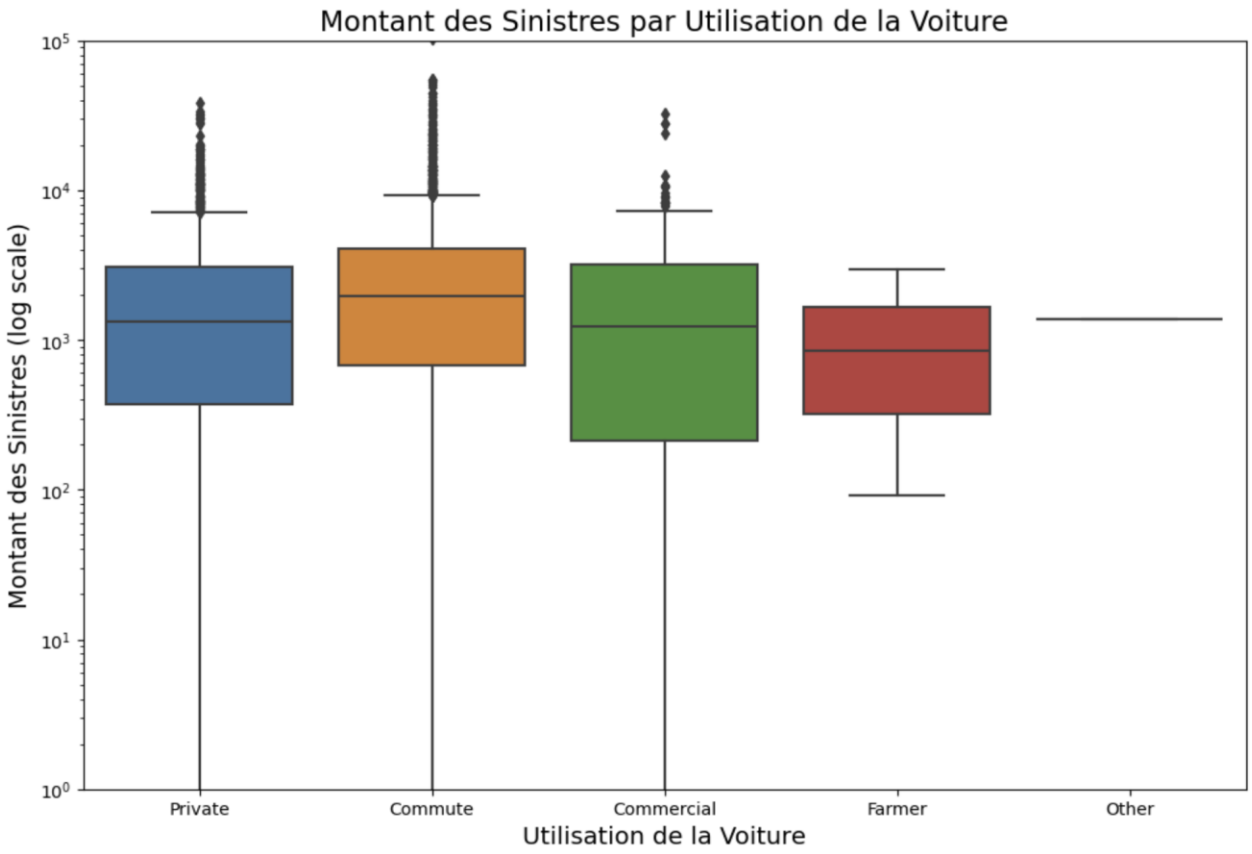
Graphique 1: Montant des Sinistres par Utilisation de la Voiture



Dans la sous section suivante, nous essayerons de capturer au mieux l'effet de l'utilisation des voitures sur le montant des sinistres en appliquant une fonction logarithmique. Cette action nous permet de standardiser et d'éviter les biais dans nos différentes variables.

4.2. Montant des Sinistres par Utilisation de la Voiture en Logarithme

Graphique 2: Montant des Sinistres par Utilisation de la Voiture en Logarithme



Dans l'analyse du montant des sinistres par utilisation de la voiture, présentée sur un graphique en échelle logarithmique, illustre la répartition du nombre de véhicules selon leur utilisation au sein de notre base de données, mettant en lumière les préférences d'usage des véhicules assurés. *La catégorie 'Commute'*, qui désigne les véhicules utilisés pour les trajets quotidiens, domine nettement, représentant la majorité des véhicules enregistrés. Cette prédominance souligne la fréquence élevée d'utilisation de ces véhicules, ce qui peut correspondre à un risque accru de sinistrés en raison de leur utilisation régulière.

Les véhicules à usage privé (*'Private'*) constituent également une part importante du total, reflétant une utilisation moins fréquente comparée aux trajets quotidiens mais toujours significative dans

l'ensemble du parc automobile. Par contraste, les véhicules commerciaux (*'Commercial'*) et agricole (*'Farmer'*) sont moins représentés, ce qui suggère une spécialisation de leur usage qui pourrait nécessiter des politiques d'assurance adaptées à des risques spécifiques moins fréquents mais potentiellement plus coûteux. La catégorie 'Other' regroupe uniquement les données que nous avons simulées après traitement dans la colonne .

Cette diversité d'usages pourrait impliquer des défis uniques en termes de tarification et de gestion des risques, offrant ainsi des opportunités pour des offres d'assurance spécialisées. La compréhension de cette distribution est cruciale pour les assureurs afin de calibrer leurs produits d'assurance pour correspondre précisément à l'exposition au risque de chaque catégorie, améliorant ainsi l'efficacité de la tarification et la satisfaction des clients tout en optimisant la gestion du risque.

L'utilisation d'une échelle logarithmique pour ce graphique révèle des détails cruciaux non apparents avec une échelle linéaire, notamment en gérant les valeurs extrêmes et en visualisant la variabilité au sein des catégories. Cette approche fournit une compréhension approfondie des liens entre l'utilisation des véhicules et les montants des sinistres, permettant aux assureurs de mieux adapter les couvertures et les primes d'assurance pour chaque segment de clients, tout en améliorant les modèles de tarification pour refléter précisément les risques réels. Par ailleurs, nous avons réalisé une analyse graphique sur la répartition de l'usage des voitures.

4.3. La répartition du nombre de véhicules selon leur utilisation

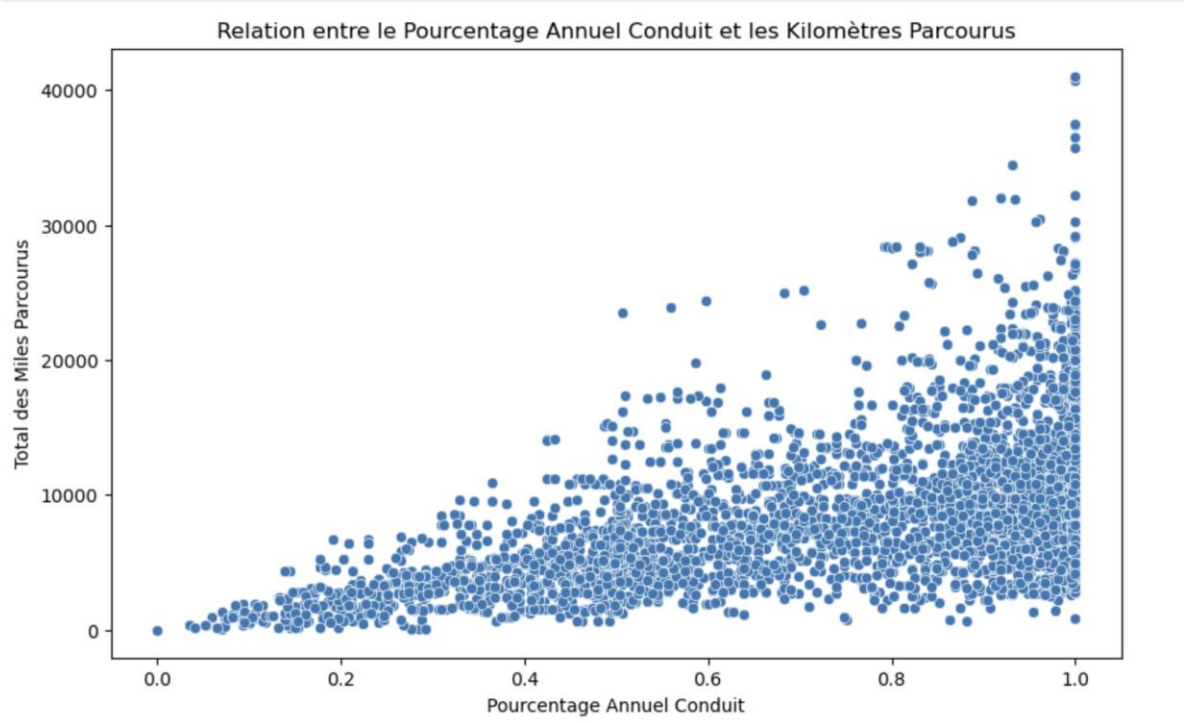
Le graphique 3 en annexe illustre la répartition du nombre de véhicules selon leur utilisation au sein de notre base de données. Les catégories d'utilisation se distinguent nettement, mettant en évidence une prédominance notable des véhicules utilisés pour les trajets quotidiens (Commute), qui représentent la majorité des véhicules analysés. Cette catégorie est suivie par les véhicules à usage privé (Private), qui constituent également une part significative du parc automobile. En revanche, les véhicules commerciaux (Commercial) et ceux utilisés à des fins agricoles (Farmer) sont nettement moins nombreux, ce qui suggère une moindre représentation dans la population étudiée.

Cette distribution a des implications directes sur la gestion des risques et la tarification des assurances. La prévalence élevée des véhicules utilisés pour les trajets quotidiens pourrait indiquer un risque plus élevé de sinistres dus à la fréquence accrue d'utilisation, tandis que la rareté relative des véhicules commerciaux et agricoles peut refléter des opportunités de niches spécialisées pour les offres d'assurance. Les assureurs peuvent tirer parti de ces informations pour développer des produits d'assurance mieux ciblés et adaptés aux spécificités de chaque catégorie d'utilisation, optimisant ainsi la couverture et les coûts pour les assurés tout en maximisant leur propre gestion du risque.

4.4. Relation entre le Pourcentage Annuel Conduit et les Kilomètres Parcourus

Le graphique ci-dessous montre une relation positive entre le pourcentage annuel conduit et les kilomètres parcourus est clairement visible, illustrant que plus un véhicule est utilisé au cours de l'année, plus il tend à parcourir de distance. Cette tendance est intuitive mais importante pour valider l'exactitude des données recueillies et peut servir à identifier des cas d'utilisation anormalement élevée ou faible qui pourraient nécessiter une enquête supplémentaire.

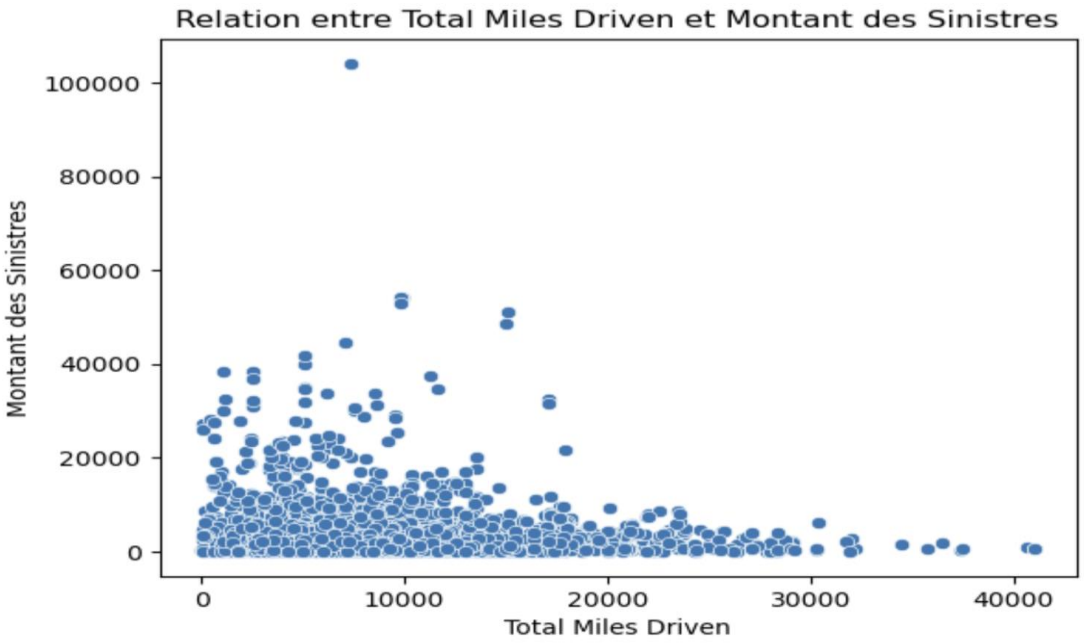
Graphique 4: Relation entre le Pourcentage Annuel Conduit et les Kilomètres Parcourus



4. 5. Relation entre Total Miles Driven et Montant des Sinistres

Ce graphique met en évidence qu'il n'y a pas une relation linéaire forte entre les miles parcourus et le montant des sinistres, suggérant que d'autres facteurs, tels que les conditions de conduite, l'environnement routier, ou le comportement au volant, pourraient jouer un rôle plus significatif dans la détermination des sinistres que simplement la distance parcourue.

Graphique 5: Relation entre Total Miles Driven et Montant des Sinistres



4.6. Montant des Sinistres en fonction de l'Âge du Véhicule

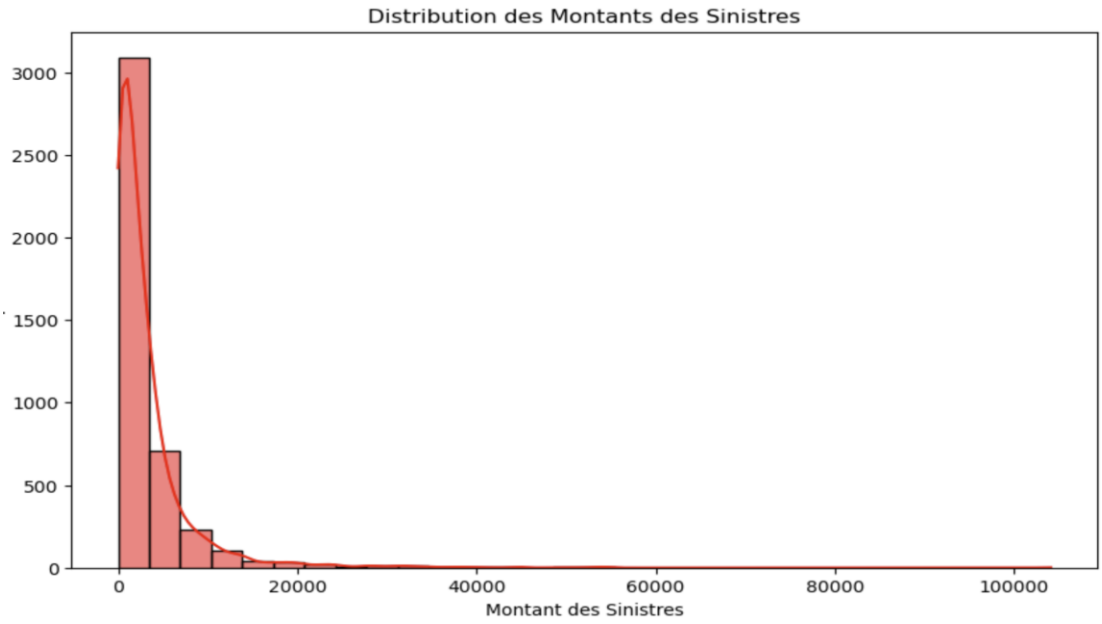
Il est intéressant de noter que l'âge du véhicule ne montre pas une tendance claire avec le montant des sinistres, indiquant que l'âge, en soi, n'est pas un prédicteur direct des coûts de sinistres. Cela peut indiquer que des facteurs tels que l'entretien du véhicule ou le type de véhicule pourraient être des indicateurs plus précis du risque de sinistre que l'âge seul (confer graphique 6 en Annexe)

4.7. Distribution des Montants des Sinistres

La distribution des montants des sinistres montre une forte concentration de sinistres de faible montant, avec quelques valeurs extrêmes beaucoup plus élevées. Cela pourrait indiquer la présence

de sinistres majeurs occasionnels qui pourraient disproportionnellement affecter la stabilité financière de l'assureur. Cela justifie l'importance d'une gestion de risque efficace et d'une tarification précise.

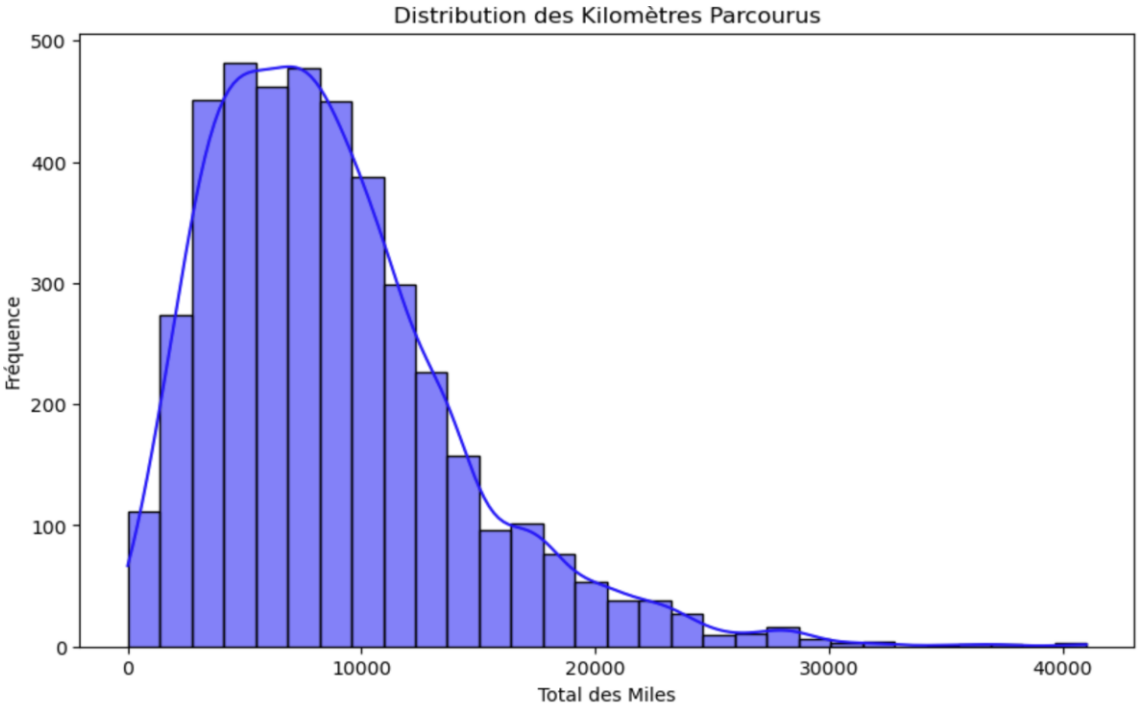
Graphique 7: Distribution des montant des sinistres



4.8. Distribution des Kilomètres Parcourus

La distribution des kilomètres parcourus montre une décroissance exponentielle, avec la majorité des véhicules parcourant relativement peu de kilomètres, tandis que quelques-uns accumulent des distances beaucoup plus grandes. Cette information est cruciale pour segmenter les assurés en groupes de risque basés sur l'utilisation du véhicule.

Graphique 8: Distribution des kilomètres parcours



5. Modélisation Statistique et de Machine Learning

Dans cette section de notre rapport, nous analysons l'impact de diverses variables sur le montant des sinistres d'assurance auto en utilisant un modèle économétrique généralisé avec une famille de distribution Gamma et une fonction de lien log.

Ce modèle révèle plusieurs connaissances clés : un temps prolongé de conduite réduit significativement le montant des sinistres, suggérant que les trajets longs sont souvent entrepris par des conducteurs plus prudents ou expérimentés. Le nombre de sinistres est positivement corrélé avec le montant des sinistres, ce qui est intuitif et aligné avec les attentes. L'âge de l'assuré et le score de crédit ont des effets respectivement positif et négatif sur les coûts des sinistres, indiquant que les jeunes conducteurs ou ceux avec un score de crédit inférieur peuvent être impliqués dans des sinistres plus coûteux.

En outre, les années sans sinistre sont associées à une réduction du montant des sinistres, validant la notion que les conducteurs avec un historique de conduite propre posent moins de risques. L'usage du véhicule pour des déplacements quotidiens ou privés est lié à des montants de sinistres plus élevés par rapport à une utilisation commerciale, soulignant les risques accrus associés aux usages fréquents.

Nos résultats peuvent guider nos décideurs dans l'ajustement des primes et le développement de stratégies préventives pour les profils à haut risque, ainsi que dans l'élaboration de produits d'assurance ciblés pour les jeunes conducteurs et ceux avec un faible score de crédit.

En conclusion, cette analyse détaillée offre des stratégies fondées sur les données pour optimiser la gestion des risques et la tarification des polices d'assurance automobile.

Tableau 2: Modèle GLM avec la loi GAMMA

Generalized Linear Model Regression Results

Dep. Variable:	AMT_Claim	No. Observations:	3412
Model:	GLM	Df Residuals:	3401
Model Family:	Gamma	Df Model:	10
Link Function:	Log	Scale:	0.14511
Method:	IRLS	Log-Likelihood:	inf
Date:	Tue, 30 Apr 2024	Deviance:	23559.
Time:	15:06:59	Pearson chi2:	494.
No. Iterations:	60	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7912	0.095	18.777	0.000	1.604	1.978
Pct.drive.2hrs	-5.4374	0.914	-5.947	0.000	-7.229	-3.645
NB_Claim	0.1586	0.030	5.366	0.000	0.101	0.217
Duration	0.0007	0.000	5.701	0.000	0.000	0.001
Insured_age	0.0025	0.001	3.215	0.001	0.001	0.004
Credit_score	-0.0005	7.7e-05	-6.717	0.000	-0.001	-0.000
Years_noclaims	-0.0040	0.001	-5.531	0.000	-0.005	-0.003
Marital_Single	0.0282	0.015	1.933	0.053	-0.000	0.057
Car_use_Commute	0.1099	0.031	3.498	0.000	0.048	0.171
Car_use_Private	0.0813	0.032	2.534	0.011	0.018	0.144
Region_Urban	-0.0138	0.017	-0.792	0.428	-0.048	0.020

5.1. Évaluation de la Performance des Modèles de Prédiction

Dans cette section, nous évaluons la performance de trois modèles statistiques différents pour prédire le montant des sinistres d'assurance auto. L'efficacité de chaque modèle a été mesurée à l'aide de plusieurs indicateurs clés : l'Erreur Quadratique Moyenne (MSE), la Racine de l'Erreur Quadratique Moyenne (RMSE), l'Erreur Absolue Moyenne (MAE), et le coefficient de détermination (R^2).

Modèle 1 affiche une MSE de 5.675, une RMSE de 2.382, une MAE de 1.665, et un R^2 de 0.102. Ces valeurs suggèrent que ce modèle a une capacité modérée à prédire avec précision le montant des sinistres, avec une variance expliquée d'environ 10.2% du total.
Modèle 2 et Modèle 3, ayant les mêmes métriques de performance, montrent une MSE de 5.800, une RMSE de 2.408, et une MAE de 1.663. Le R^2 pour ces modèles est de 0.082, indiquant que moins de 8.3% de la variance des montants des sinistres est expliquée par ces modèles par rapport au premier modèle.

5.2. Optimisation du Modèle de Prédiction des Montants des Sinistres

Le développement initial de notre Modèle 1 a révélé la présence de plusieurs variables non significatives, ce qui a impacté négativement la performance de la prédiction. En conséquence, une réévaluation des variables incluses a été effectuée, aboutissant à l'élaboration du Modèle 3. Ce dernier modèle a été affiné en retirant les variables qui ne contribuaient pas significativement à la prédiction des montants des sinistres, dans le but d'améliorer la précision et l'efficacité du modèle.

Le modèle 1 initial comportait des variables supplémentaires qui, bien que potentiellement intéressantes, ont dilué la capacité prédictive du modèle, comme indiqué par un R^2 de 0.102. Après élimination de ces variables non significatives, nous avons développé le Modèle 3, qui, bien que similaire en performance avec le Modèle 2, présente une structure plus simplifiée et potentiellement plus robuste en termes d'interprétabilité et de généralisation.
Les métriques de performance du Modèle 3 — MSE de 5.800, RMSE de 2.408, MAE de 1.663, et un R^2 de 0.082 — confirment que bien que l'amélioration en termes de capacité explicative soit marginale, le retrait des variables non pertinentes est une étape cruciale dans l'optimisation de modèles économétriques. Cette approche assure non seulement une meilleure compréhension des

facteurs influençant les sinistres, mais favorise également une modélisation plus efficace en évitant la suradaptation et en améliorant la généralisabilité du modèle.

La réduction des variables a également des implications pratiques pour les stratégies d'assurance, car elle permet aux assureurs de se concentrer sur les facteurs les plus impactants, réduisant ainsi les coûts associés à la collecte et à l'analyse de données superflues. Ce processus d'affinage souligne l'importance de la sélection de variables dans la modélisation prédictive et incite à une exploration continue pour une compréhension encore plus approfondie des dynamiques sous-jacentes des sinistres.

5.3. Modélisation Prédictive et Analyse des Facteurs Influents sur les Montants des Sinistres d'Assurance

Dans le cadre de notre étude sur l'impact des différentes variables sur le montant des sinistres d'assurance, nous avons appliqué des techniques avancées de modélisation statistique et de machine learning pour prédire les montants des sinistres. L'analyse a débuté par l'utilisation d'un modèle de régression linéaire généralisée (GLM) avec une distribution Gamma et une fonction de lien logarithmique, révélant des influences significatives de plusieurs variables, notamment l'usage de la voiture, l'âge assuré, et le nombre d'années sans sinistre.

Pour affiner notre compréhension des dynamiques sous-jacentes et améliorer la précision des prédictions, des modèles de machine learning tels que les arbres de décision et la forêt aléatoire (Random Forest) ont été employés. Les résultats de l'arbre de décision ont montré une importance marquée de variables telles que le score de crédit, le pourcentage annuel de conduite, et les comportements de conduite spécifiques tels que le freinage à différentes intensités. Le modèle Random Forest, en particulier, a offert une amélioration substantielle de la précision avec un R-squared de 0.5579, indiquant une capacité modérée à expliquer la variance des montants des sinistres par les variables incluses.

L'analyse de l'importance des caractéristiques a révélé que certains comportements de conduite, tels que l'utilisation de la voiture à des fins commerciales et les habitudes de freinage, sont particulièrement prédictifs des montants des sinistres. Ces insights sont cruciaux pour les stratégies

de tarification et de gestion des risques, soulignant la nécessité de tarifs différenciés et de mesures incitatives pour encourager des comportements de conduite plus sûrs.

En conclusion, les modèles développés permettent non seulement de prédire de manière plus précise les montants des sinistres, mais offrent également des perspectives pour une gestion plus ciblée des politiques d'assurance. Ces résultats soutiennent la mise en place de stratégies personnalisées qui pourraient potentiellement réduire les coûts pour les assureurs et les assurés, tout en améliorant la sécurité routière.

5.4. Comparaison des Performances des Modèles de Prédiction des Sinistres

La table suivante fournit un comparatif des variables clés et des performances de trois modèles statistiques avancés : le modèle linéaire généralisé (GLM), le Random Forest et le XGBoost. Cette comparaison vise à identifier les modèles les plus efficaces pour prédire le montant des sinistres en se basant sur divers indicateurs de performance tels que l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne (MAE) et le coefficient de détermination (R^2). L'analyse révèle les forces et les limites de chaque modèle en fonction de la précision et de la fiabilité des prédictions obtenues.

Variables Clés	GLM	Random Forest	XGBoost
Brake.08miles	X		X
Pct.drive.2hrs	X	X	
NB_Claim	X		X
Insured_age	X		X
Total.miles.driven		X	X
MSE (en millions)	5.80	9.06	9.49
MAE	1.66	1.90	2.02
R^2	0.08	0.56	0.54

Notre étude comparative des modèles GLM (Generalized Linear Model), Random Forest et XGBoost révèle des variations significatives dans leur capacité à prédire les montants des sinistres, chacun mettant en avant différentes variables clés. Le GLM présente le plus faible Mean Squared Error, indiquant une précision accrue dans la minimisation des erreurs de prédiction, bien que son coefficient de détermination

R^2 reste relativement bas, ce qui souligne une capacité limitée à expliquer la variance des montants des sinistres.

D'autre part, le modèle Random Forest se distingue par le meilleur R^2, confirmant son efficacité à capter une plus grande partie de la variabilité des montants des sinistres. Le modèle XGBoost, tout en ayant un MSE comparable à celui de Random Forest, montre un R^2 légèrement inférieur, ce qui pourrait indiquer une moindre efficacité pour certaines distributions de données.

En ce qui concerne les variables prédictrices, Brake.08miles, Pct.drive.2hrs, et NB_Claim figurent parmi les plus influentes dans plusieurs modèles, ce qui met en lumière leur rôle crucial dans l'évaluation des risques et des montants de sinistres. notamment, la variable Insured_age est significative dans les modèles GLM et XGBoost, mais pas dans Random Forest, suggérant que les modèles basés sur les arbres et les modèles linéaires peuvent différer dans leur traitement des facteurs démographiques.

Enfin, Total.miles.driven se révèle être un prédicteur clé pour les modèles basés sur les arbres, soulignant l'importance de l'utilisation du véhicule dans l'estimation des sinistres. Ces insights suggèrent qu'une stratégie de modélisation combinée pourrait être bénéfique pour tirer pleinement parti des forces de chaque approche. Une analyse plus fine de ces variables et l'intégration de plusieurs modèles pourraient ainsi contribuer à affiner les prédictions des montants des sinistres, offrant une base solide pour des stratégies de tarification et de gestion des risques plus précises et efficaces.

6. Limites et Suggestions pour Améliorations Futures

Bien que notre modèle offre des aperçus significatifs sur les prédictifs des montants des sinistres, plusieurs axes d'amélioration restent envisageables pour optimiser la pertinence et la précision des prédictions futures. Une intégration de données contextuelles plus détaillées, telles que les conditions climatiques ou les infrastructures routières, pourrait enrichir l'analyse et permettre une meilleure compréhension des facteurs influençant les sinistres. Il est également essentiel de

considérer les interactions complexes entre les variables, qui peuvent parfois être masquées dans les modèles simplifiés.

La prudence est de mise dans l'interprétation des résultats, surtout en raison des transformations appliquées à la variable de réponse et des limites inhérentes aux modèles statistiques utilisés. Pour pallier ces limites, il serait judicieux de conduire des études complémentaires, notamment des analyses de sensibilité et des validations croisées, pour confirmer la robustesse des modèles.

Ce modèle représente une avancée importante vers une compréhension plus fine des dynamiques de sinistralité dans le secteur de l'assurance automobile. Avec des validations et des ajustements supplémentaires, il pourrait servir de fondement au développement d'outils prédictifs opérationnels.

7. Recommandations pour la Pratique.

Sur la base des analyses effectuées, nous recommandons l'adoption de politiques de tarification dynamique, prenant en compte les variables clés identifiées. Par ailleurs, des initiatives telles que des programmes de conduite défensive pour les jeunes conducteurs et des stratégies de gestion financière pourraient contribuer à atténuer les risques et à réduire la fréquence des sinistres.

Les pratiques de conduite sécuritaire et une bonne gestion financière, comme en témoignent les périodes prolongées sans réclamations et les bons scores de crédit, devraient être encouragées. Ces mesures pourraient se traduire par des réductions de primes pour les conducteurs exemplaires. En outre, l'âge et le statut marital, influençant significativement les montants des sinistres, pourraient justifier une segmentation plus affinée des offres d'assurance, permettant ainsi de mieux gérer le risque selon les profils.

8. Conclusion et Perspectives

Ce projet a mis en lumière les facteurs clés qui influencent les montants des sinistres dans l'industrie de l'assurance automobile, en utilisant des techniques avancées d'analyse statistique et de machine learning. Nos analyses ont révélé que des variables telles que l'utilisation de la voiture, les pratiques de conduite, l'âge de l'assuré, et le score de crédit jouent un rôle prépondérant dans la prédiction des montants des sinistres.

En particulier, nous avons découvert que les conducteurs qui utilisent leur véhicule principalement pour des trajets quotidiens ou à des fins commerciales tendent à déclarer des sinistres plus élevés, ce qui suggère une corrélation entre l'intensité de l'utilisation de la voiture et la fréquence des accidents. De même, les analyses ont montré que des périodes prolongées sans réclamation et un bon score de crédit sont associés à des montants de sinistres plus faibles, soulignant l'importance des comportements de conduite prudente et de la stabilité financière.

Cependant, malgré les insights significatifs obtenus, notre étude reconnaît certaines limites, notamment la nécessité d'intégrer davantage de données contextuelles telles que les conditions météorologiques et les conditions de route, qui pourraient affecter les montants des sinistres.

Sur la base de ces constatations, nous recommandons l'adoption de politiques de tarification dynamique qui tiennent compte des variables identifiées. Des initiatives telles que les programmes de conduite défensive pour les jeunes conducteurs pourraient non seulement aider à réduire la fréquence des sinistres mais aussi diminuer leur gravité.

Pour finir, ce travail offre une base solide pour le développement futur d'outils prédictifs plus précis et personnalisés en assurance automobile. L'avenir de la modélisation des risques dans l'assurtech semble prometteur, avec des possibilités d'intégrer des analyses en temps réel et d'exploiter les technologies émergentes pour une compréhension encore plus profonde et une gestion efficace des sinistres. Ce projet n'est qu'un début dans l'amélioration continue de nos capacités prédictives et de notre compréhension des dynamiques sous-jacentes à l'assurance automobile.

Annexes

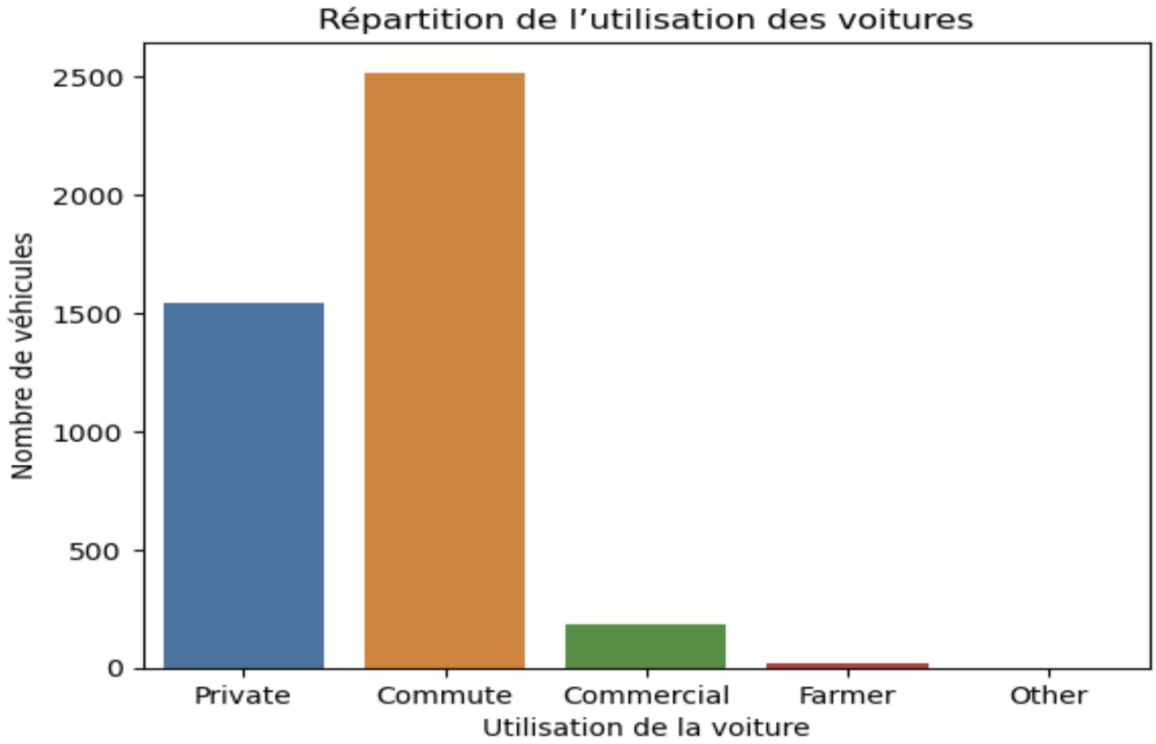
Tableau 1: Description de variables

Nom	Description
Id_pol	Identifiant unique de la police d'assurance.
Annual.pct.driven	Pourcentage de l'année pendant lequel le véhicule est conduit.
Total.miles.driven	Nombre total de miles parcourus pendant l'année.
Pct.drive.mon à Pct.drive.sun	Pourcentage du total des miles parcourus chaque jour de la semaine (lundi à dimanche).
Pct.drive.2hrs, Pct.drive.3hrs, Pct.drive.4hrs	Pourcentage du total des miles parcourus en sessions de conduite de 2 heures, 3 heures et 4 heures, respectivement.
Pct.drive.wkday et Pct.drive.wkend	Pourcentage du total des miles parcourus en semaine (jours ouvrables) et le weekend.
Pct.drive.rush am et Pct.drive.rush pm	Pourcentage du total des miles parcourus pendant les heures de pointe du matin et du soir.
Avgdays.week	Nombre moyen de jours par semaine pendant lesquels le véhicule est conduit.
Accel.06miles à Accel.14miles	Nombre d'accélérations fortes par 100 miles, classées par intensité (par exemple, 0.6 miles par heure^2 à 1.4 miles par heure^2).
Brake.06miles à Brake.14miles	Nombre de freinages forts par 100 miles, également classés par intensité.
Left.turn.intensity08 à Left.turn.intensity12 et Right.turn.intensity08 à Right.turn.intensity12	Mesures de l'intensité des virages à gauche et à droite, probablement notées sur une échelle ou selon des seuils spécifiques d'intensité de virage.
Duration	Durée de la police d'assurance.
Insured_age	Âge de l'assuré.
Insured_sex	Sexe de l'assuré.
Car_age	Âge du véhicule assuré.
Marital	Statut marital de l'assuré.
Car_use	Utilisation du véhicule (personnel, professionnel, etc.).
Credit_score	Score de crédit de l'assuré.
Region	Région géographique.
Annual_miles_drive	Nombre annuel de miles parcourus.
Years_noclaims	Années sans sinistres.
Territory	Territoire géographique.
AMT_Claim	Montant des sinistres.
NB_Claim	Nombre de sinistres.

Tableau 2: Tableau de Statistiques Descriptives des Variables de Conduite

Variable	Min	25%	Médiane	Moyenne	75%	Max	Ecart-type
Annual.pct.driven	0.00%	52.88%	83.56%	74.56%	97.26%	100.00%	25.07%
Total.miles.driven	0	4,781.47	7,771.44	8,733.35	11,359.42	41,019.58	5,440.23
Pct.drive.mon	0.00%	12.57%	13.95%	14.02%	15.29%	31.34%	2.45%
Pct.drive.tue	0.00%	13.41%	14.75%	14.95%	16.35%	40.21%	2.71%
Pct.drive.wed	0.00%	13.34%	14.83%	14.84%	16.22%	35.51%	2.57%
Pct.drive.thr	0.00%	14.19%	15.46%	15.58%	16.89%	49.80%	2.60%
Pct.drive.fri	0.00%	14.30%	15.61%	15.74%	17.01%	39.97%	2.58%
Pct.drive.sat	0.00%	11.56%	13.49%	13.73%	15.59%	54.21%	3.78%
Pct_drive_sun	0.00%	9.11%	11.14%	11.12%	13.03%	31.63%	3.40%
Avgdays.week	0.00	5.40	6.04	5.83	6.49	7.00	0.90
Accel.06miles	0	13	30	50.95	64	621	61.80
Brake.06miles	0	42	75	99.09	129	621	82.38
AMT_Claim	$0	$518.82	$1,708.10	$3,179.15	$3,735.48	$104,074.89	$5,163.25
Insured_age	18	35	47	46.84	58	90	14.52
Car_age	0	1	4	4.55	7	18	3.59
Credit_score	428	715	792	769.84	838	900	91.18
Annual_miles_drive	683.51	6,213.71	9,320.57	9,915.03	12,427.42	33,802.58	4,007.57
Years_noclaims	0	10	23	23.83	35	74	15.04

Graphique 3 : Répartition de l'utilisation des voitures



Graphique 6 : Relation entre le montant des sinistres et l'âge des véhicules