Rapport Projet Data

Partie Machine Learning

Nom et Prénoms : Amégadjaka kossi Clément

Sommaire :

1- Introduction

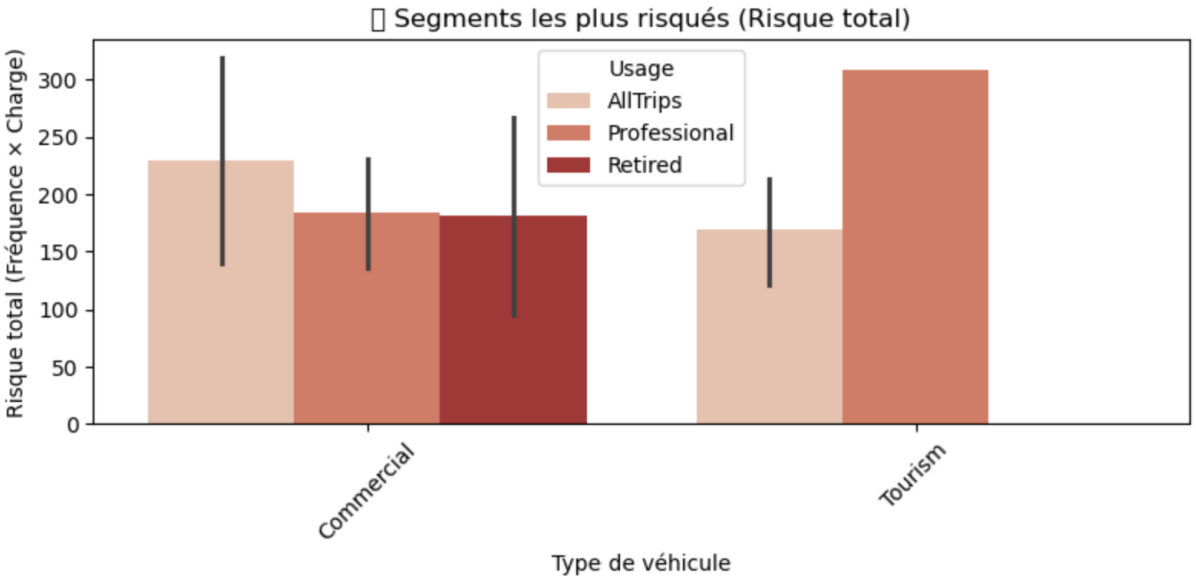
Dans un contexte où la donnée constitue un levier stratégique essentiel, j'ai été sollicité en tant que data scientists pour explorer de nouvelles bases d'assurance au sein de notre insurtech. L'objectif est d'enrichir notre compréhension des comportements liés aux sinistres en intégrant des dimensions encore peu exploitées, comme l'impact climatique, les caractéristiques techniques des véhicules et les profils des assurés. Au moyen de Python, j'ai mené un projet organisé autour de trois grands volets : la création d'un véhiculier destiné à repérer les véhicules présentant un risque accru, en croisant les données techniques automobiles avec les historiques de sinistres ; l'analyse spatiale du territoire français, avec la construction d'un zonier climatique fondé sur les codes INSEE et les indicateurs météorologiques départementaux ; la modélisation statistique de la sinistralité, mobilisant plusieurs méthodes complémentaires : une régression logistique pour évaluer l'effet des zones climatiques sur la probabilité de survenue d'un sinistre (défini par claim_amount > 0) ; un GLM (Generalized Linear Model) pour mesurer l'impact des différentes zones sur le montant des sinistres ; un arbre de décision pour prédire la survenance d'un sinistre à partir des données techniques et géographiques. La démarche adoptée a combiné un travail approfondi de préparation des données (nettoyage, enrichissement, jointures) avec des analyses statistiques et des modèles prédictifs, afin d'identifier les facteurs influençant la fréquence et le coût des sinistres.

2- Préparation de données

Le projet s'appuie sur trois principales sources de données croisées : Données assurantielles et de sinistralité (pg17trainpol et pg17trainclaim) : elles décrivent le profil des assurés, les caractéristiques de leurs contrats et véhicules, ainsi que les sinistres déclarés (montant et nombre). Les bases ont été préparées pour distinguer les assurés avec et sans sinistre.

3- Construction du véhiculier.

Dans la continuité de notre analyse, j'ai entamé une démarche de machine learning en exploitant la base totale des assurés, incluant à la fois les sinistrés et les non-sinistrés, afin de construire un véhiculier destiné à identifier les profils de véhicules présentant un risque accru.



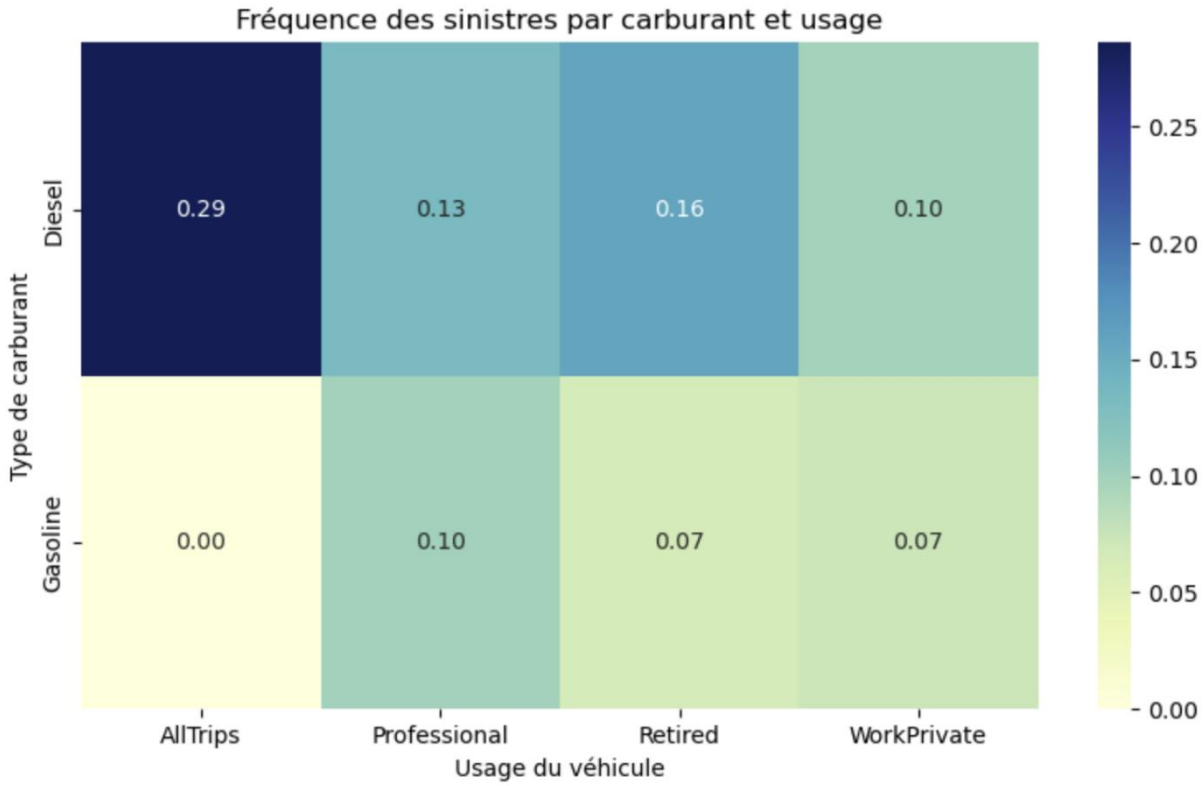
Interprétation :

L'analyse du risque total (défini comme le produit de la fréquence des sinistres par leur charge moyenne) met en évidence des disparités significatives selon le type de véhicule et son usage déclaré. Les véhicules de type Tourisme utilisés dans un contexte professionnel se distinguent comme le segment le plus risqué, affichant des niveaux de sinistralité supérieurs aux autres catégories. Cette situation peut être attribuée à une exposition accrue aux risques routiers liée à une utilisation intensive, souvent dans un cadre professionnel nécessitant de nombreux déplacements. Les véhicules commerciaux utilisés pour tous types de trajets ou à des fins professionnelles présentent également un risque non négligeable, bien que globalement inférieur à celui des Tourismes professionnels. Enfin, les véhicules associés à un usage retraité affichent un risque modéré, probablement en lien avec une fréquence d'utilisation plus faible et des déplacements moins exposés.

Ces résultats suggèrent que des stratégies différenciées de tarification et de prévention doivent être mises en œuvre, notamment en renforçant la surveillance et la gestion du risque sur le segment des véhicules de tourisme à usage professionnel.

4- Fréquence des sinistres : influence du type de carburant et du profil d'utilisation

Afin de mieux comprendre l'impact du type de carburant et de l'usage des véhicules sur la sinistralité, une analyse croisée a été réalisée. Cette étude vise à identifier les combinaisons présentant des fréquences de sinistre plus élevées, dans une optique d'ajustement des politiques de tarification et de prévention.



Interprétation :

L'analyse révèle que les véhicules diesel utilisés pour tous types de trajets (AllTrips) présentent la fréquence de sinistres la plus élevée (29 %). Cela suggère que ces véhicules, probablement soumis à un usage intensif et diversifié, sont plus exposés aux risques. En comparaison, les véhicules diesel à usage professionnel ou retraité enregistrent des fréquences de sinistres modérément élevées (respectivement 13 % et 16 %), tandis que l'usage privé des diesels affiche une fréquence plus faible (10 %). En ce qui concerne les véhicules essence (Gasoline), les fréquences de sinistres sont globalement plus basses quel que soit l'usage, oscillant entre 0 % pour AllTrips et 10 % pour Professionel, avec des valeurs plus faibles encore pour les usages Retired et WorkPrivate.

Ainsi, la sinistralité semble plus marquée chez les véhicules diesel que chez les essences (Gasoline), et particulièrement pour des usages intensifs et variés.

Ces résultats confortent l'idée que la combinaison type de carburant et mode d'utilisation est un facteur déterminant du risque, et devrait être intégrée dans les modèles de scoring assurantiels pour affiner les politiques tarifaires.

5- Régression logistique pour analyser l'effet de la zone climatique sur la probabilité de sinistre (claim_amount > 0)

Cette régression vise à comprendre l'influence des différentes zones climatiques sur la probabilité qu'un assuré déclare un sinistre. La France a été découpée en quatre grandes zones

climatiques, allant du climat méditerranéen chaud (zone 1) au climat froid de montagne (zone 3).

Variables	Coefficients	P-Value
Constante	-2.0592	0.000
C(zone_climatique)[T.2.0]	-0.0490	0.155
C(zone_climatique)[T.3.0]	-0.0572	0.008

Interprétation

Pour la suite de l'analyse des variables assurantielles et véhiculaires, une régression logistique a été menée afin d'évaluer l'effet de la zone climatique de résidence sur la probabilité de déclaration d'un sinistre. Pour cela, les départements français ont été regroupés en quatre grandes zones climatiques distinctes :

Zone 0 : climat océanique humide (Bretagne, Normandie)

Zone 1 : climat méditerranéen chaud (Sud-Est)

Zone 2 : climat continental tempéré (Centre, Bourgogne)

Zone 3 : climat froid et d'altitude (Massif central, Alpes, Pyrénées)

Les résultats montrent que le climat a un impact statistiquement significatif mais modéré sur la sinistralité. Plus précisément, les assurés résidant dans les zones de climat froid (zone 3) présentent une probabilité de sinistre légèrement inférieure à ceux situés dans les régions humides de l'ouest (zone 0). Ce constat pourrait s'expliquer par des comportements de conduite plus prudents dans des environnements routiers réputés difficiles (neige, verglas). En revanche, aucune différence significative n'a été observée pour la zone continentale tempérée (zone 2) par rapport à la zone de référence.

Toutefois, l'effet global de la zone climatique reste faible au regard du pseudo R^2 très bas du modèle. Cela confirme que si l'environnement géographique peut influencer marginalement la sinistralité, les principaux facteurs explicatifs résident davantage dans les caractéristiques individuelles des conducteurs, des véhicules et des contrats.

6- Modélisation du montant des sinistres selon la zone climatique

Dans le prolongement de l'analyse de la probabilité de sinistre, j'ai construit un modèle linéaire généralisé (GLM) afin d'évaluer l'impact des différentes zones climatiques sur le montant des sinistres déclarés par les assurés. Cette approche vise à comprendre si certaines conditions climatiques sont associées à des coûts moyens de sinistres plus élevés ou plus faibles.

Variables	Coefficients	P-value
Constante	6.9639	0.000
C(zone_climatique)[T.2.0]	-0.0655	0.416
C(zone_climatique)[T.3.0]	-0.1167	0.019

Interprétation :

Afin d'approfondir l'impact du climat sur les coûts de sinistres, un modèle de régression linéaire généralisée (GLM) a été mis en œuvre. L'objectif était de tester si l'appartenance à une zone climatique influençait de manière significative le montant des sinistres déclarés. Le modèle repose sur une famille Gamma avec une fonction de lien logarithmique, choix adapté aux montants strictement positifs et asymétriques. Il a été estimé à partir de 10 981 sinistres.

Les principaux résultats sont les suivants :

Le pseudo R^2 du modèle est de 0,0005, indiquant que les seules zones climatiques expliquent peu de la variabilité totale des montants de sinistres. Toutefois, ce niveau est cohérent compte tenu de la nature volatile de cette variable dans le secteur assurantiel. La zone 2 (climat continental tempéré : Bourgogne, Centre-Val de Loire) n'affiche pas d'effet significatif sur les montants, par rapport à la zone 0 (climat océanique humide : Bretagne, Normandie). Cela se traduit par une p-value de 0,413. En revanche, la zone 3 (climat froid et montagneux : Massif Central, Alpes, Pyrénées) présente un effet significatif, avec un coefficient négatif (-0,1167, p-value = 0,019). Cela signifie que les sinistres y sont en moyenne moins coûteux que dans les régions de climat océanique.

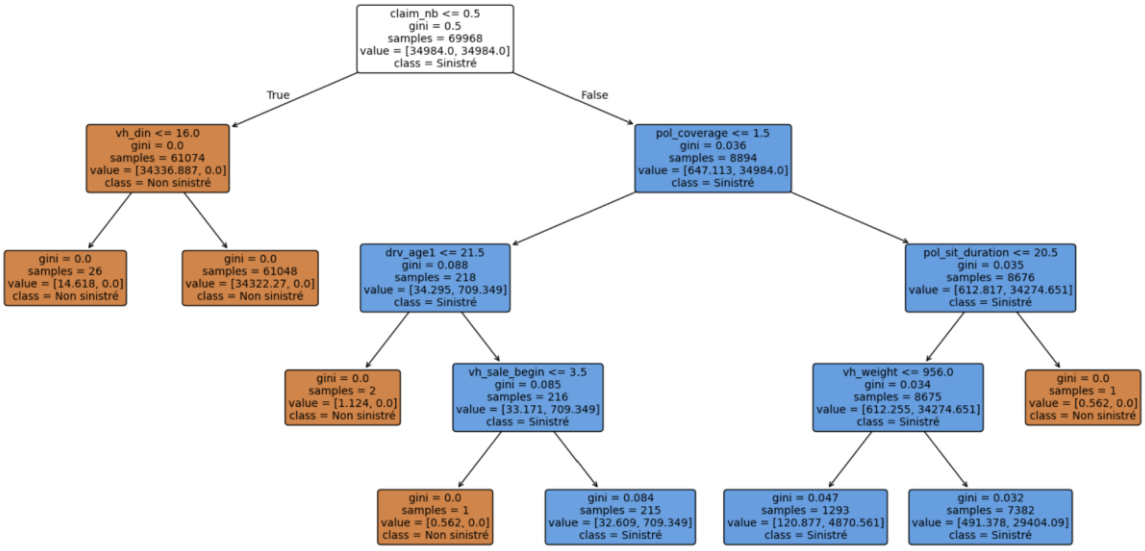
Les assurés situés dans les zones climatiques froides (zone 3) génèrent des sinistres financièrement moins lourds que ceux situés dans les zones au climat plus tempéré. Plusieurs hypothèses peuvent être avancées telles que :

Une conduite plus prudente liée aux conditions météorologiques difficiles (neige, gel, verglas) ; des trajets plus courts et un usage modéré du véhicule en raison de la géographie montagneuse ou encore un parc automobile potentiellement différent (véhicules plus adaptés ou moins performants, donc causant des dommages moindres).

Ces éléments soulignent l'intérêt d'intégrer la zone climatique comme une variable complémentaire dans les modèles de tarification et de segmentation du portefeuille clients.

7- Arbre de décision pour prédire les sinistres

L'objectif de cette modélisation est d'utiliser un arbre de décision pour estimer la probabilité qu'un assuré déclare un sinistre (classe 1) ou reste sans sinistre (classe 0), en se basant sur les caractéristiques de son contrat, de son véhicule et de son profil. L'arbre permet de mettre en évidence les variables explicatives les plus déterminantes, tout en offrant une interprétation simple et visuelle des règles de décision.



Interprétation :

L'arbre de décision construit pour prédire la déclaration de sinistre met en lumière plusieurs variables clés. La première scission se fait sur le nombre de sinistres déjà déclarés (claim_nb). Logiquement, les assurés ayant un historique sans sinistre (claim_nb $\leq 0{,}5$) sont d'abord orientés vers la classe "non sinistré". Parmi eux, la cylindrée du véhicule (vh_din) joue un rôle discriminant supplémentaire : les véhicules avec une faible puissance (vh_din ≤ 16) renforcent la probabilité de non-sinistralité.

Pour les assurés ayant déjà déclaré au moins un sinistre (claim_nb > 0,5), d'autres variables entrent en jeu. La couverture d'assurance (pol_coverage) est un facteur important : une faible couverture est associée à un risque de sinistre plus élevé. D'autres critères comme la durée de détention du contrat (pol_sit_duration), le poids du véhicule (vh_weight), l'âge du conducteur (drv_age1) ou l'ancienneté de mise en circulation du véhicule (vh_sale_begin) permettent de raffiner davantage la prédiction.

L'évaluation de l'arbre sur l'échantillon test donne d'excellents résultats :

L'Accuracy (taux de bonne classification) : 98 %, traduisant une excellente capacité de l'arbre à distinguer sinistrés et non sinistrés. La précision sur les non-sinistrés (classe 0) : 100 % (seulement 483 faux positifs sur 26 715 non sinistrés). Le rappel sur les sinistrés (classe 1) : 100 %, ce qui signifie que tous les assurés ayant réellement eu un sinistre ont été correctement détectés par le modèle. La précision sur les sinistrés : 87 %, indiquant que quelques assurés ont été prédits à tort comme sinistrés. Le score F1 : 0,93 pour les sinistrés, confirmant un très bon équilibre entre précision et rappel.

Malgré la très bonne performance obtenue avec l'arbre de décision, certaines limites subsistent tels que la détection de 483 faux positifs, et de gestion fine des erreurs de classification. Afin d'améliorer encore la robustesse et la précision du modèle, il est judicieux d'explorer une

approche par gradient boosting, et plus particulièrement le modèle XGBoost, reconnu pour son efficacité sur des problématiques de scoring assurantiel.

8- Analyse plus fine par segment : XGBoost pour prédire les sinistres.

Vrais négatifs : 26236	Faux négatifs : 33
Faux positifs : 479	Vrais positifs : 3239

Le modèle XGBoost appliqué à la prédiction de sinistres montre une excellente performance globale avec une accuracy de 98 % sur l'échantillon de test.

Vrais négatifs (VN) : 26 236 assurés non sinistrés ont été correctement classés comme tels.

Faux positifs (FP) : 479 assurés non sinistrés ont été incorrectement prédits comme sinistrés. Cela reste un volume faible, indiquant un faible taux d'alerte injustifiée.

Faux négatifs (FN) : 33 assurés sinistrés ont été incorrectement prédits comme non sinistrés. Ce nombre est très limité, ce qui est crucial car les faux négatifs sont coûteux pour un assureur (un sinistre non anticipé).

Vrais positifs (VP) : 3239 assurés sinistrés ont été correctement identifiés comme sinistrés.

Précision sur les sinistrés (classe 1) : 87 %, indiquant qu'une grande majorité des assurés prédits comme "à risque" le sont réellement.

Recall sur les sinistrés (classe 1) : 99 %, ce qui montre que presque tous les sinistrés sont détectés par le modèle (seulement 1 % échappent).

F1-score sur les sinistrés : 93 %, équilibrant efficacement précision et rappel.

Le modèle présente un excellent compromis entre détection des sinistrés et limitation des erreurs d'alerte. Le faible nombre de faux négatifs (FN) est particulièrement rassurant dans un contexte de gestion du risque assurantiel où il est prioritaire de ne pas sous-estimer les risques.

Bien que le modèle présente de très bonnes performances, la présence de faux positifs, qui peuvent engendrer des coûts inutiles pour l'entreprise, justifie de poursuivre l'optimisation des hyperparamètres afin d'obtenir un modèle encore plus robuste.

Analyse de la matrice de confusion après optimisation des hyperparamètres

Vrais négatifs : 26231	Faux négatifs : 1
Faux positifs : 484	Vrais positifs : 3271

Le risque de faux négatif (passer à côté d'un sinistre) est quasi nul (seulement 1 erreur). Le nombre de faux positifs reste faible (484), mais comme prédire à tort qu'un client est à risque

peut entraîner des actions inutiles (hausse tarifaire, exclusion). Le poids du modèle reste largement favorable pour la détection correcte des sinistrés.

9- Conclusion

À travers cette étude, l'application des méthodes de machine learning a permis d'apporter un éclairage stratégique sur les mécanismes de sinistralité automobile. L'utilisation combinée de modèles de classification (régression logistique, arbre de décision, XGBoost) et de régression (GLM) a permis d'identifier les facteurs clés influençant la probabilité et le coût des sinistres. L'intégration de variables exogènes, notamment climatiques via la construction d'un zonier ACP, a enrichi l'analyse prédictive et mis en évidence l'impact du contexte géographique sur la sinistralité. Les arbres de décision et XGBoost ont démontré des performances robustes en termes de précision et de rappel, bien qu'une attention particulière doive être portée aux faux positifs, coûteux en gestion pour l'assureur. Par ailleurs, l'étude a permis de construire un premier véhiculier de risque, croisant les caractéristiques des véhicules et leurs usages, et de segmenter la population des assurés non sinistrés via du clustering (CAH), offrant ainsi des pistes d'actions commerciales ciblées et de prévention.

Globalement, ce travail renforce la capacité de l'entreprise à mieux tarifer, mieux cibler et mieux prévenir les risques, tout en soulignant l'importance d'une exploitation intelligente et croisée des données assurantielles, climatiques et comportementales.