Institut de Statistique de Sorbonne Université

Tarification Auto au Brésil

Groupe:
Jérémy XU
Miguel Angel BAENA BLANDON
Toluwalokpe DAKPOGAN

Professeur:
Arnaud COHEN
Kezhan SHI

Rapport du projet de Machine Learning

January 4, 2025

Table des matières

1 Présentation générale du projet et des données

1.1 Présentation de la problématique

Aujourd'hui la majorité des assureurs basent leurs tarifs sur des analyses GLM, les modèles de machine learning étant globalement peu déployés. Pourtant, ces méthodes s'avèrent souvent pertinentes, voire parfois plus performantes a priori que les approches classiques. Il sera intéressant à l'avenir de tester les deux familles d'approches lors des revues des tarifaires et de déterminer au cas par cas celle qui est la plus pertinente, en mettant en perspective les gains techniques espérés et les coûts issus de l'application des nouvelles méthodes.

1.1.1 Problématique :

Comment intégrer efficacement les méthodes de machine learning dans les revues tarifaires des assureurs, afin d'optimiser la précision des modèles tout en équilibrant les gains techniques espérés et les coûts d'implémentation par rapport aux approches traditionnelles basées sur les GLM ?

1.1.2 Tarification auto

Définition de la Prime

- La prime pure correspond au coût du risque supporté par l'assureur, soit l'espérance des sinistres.
- Elle reflète le montant attendu des indemnisations en cas de sinistre.
- **Prime finale** = Prime pure + Chargements :
  - **Chargement technique** : Protection contre la volatilité du risque (fréquence, coûts).
  - **Chargements commerciaux** : Frais d'acquisition et de gestion.
- La prime pure correspond à l'espérance d'une variable aléatoire, représentant la charge totale des sinistres survenus au cours de l'exercice.

Approche de Modélisation

Approche collective : Analyse du portefeuille dans son ensemble, contrairement à l'approche individuelle (contrat par contrat).

Formule : La charge totale des sinistres s'écrit :

$$S = \sum_{i=1}^{N} X_i$$

où :

- N : Variable aléatoire, nombre de sinistres déclarés.
- X_i : Coût du i-ème sinistre, variables indépendantes et identiquement distribuées.

Hypothèses :

- Fréquence (N) et coût (X_i) des sinistres sont indépendants.

Modéliser la prime pure revient à estimer :

- Le nombre moyen de sinistres.
- Le coût moyen des sinistres.

1.2 Présentation des données

1.2.1 Structure générale

La base de données est issue du site de données gouvernementales brésilien, similaire à data.gouv en France.Création de notre base de données en allant chercher les différentes table de description dans le site web. Cette base, divisée en une base assurée R AUTO.csv et une base de sinistres S AUTO.csv, compile d'abord des informations sur l'assuré, telles que sa prime (fournie par un assureur anonymisé), sa date de naissance, son véhicule et l'année de son véhicule , son type de couverture , le montant total de l'indemnité versé si sinistre. Ensuite, dans la base des sinistres, on trouve les occurrences de sinistres et le montant des indemnités associées, en utilisant le code de l'assuré pour faire la liaison avec la base assurée.

1.2.2 Présentation des variables

Catégorie	Variable	Description
Informations liées aux accidents	REGIAO	Localisation géographique
	CEP	Code postal
	EVENTO	Type de sinistre
	CAUSA	Cause spécifique
Informations liées aux personnes	SEXO	Sexe de l'assuré
	AGE	Âge de l'assuré
	DATA_NASC	Date de naissance
	UTILIZACAO	Type d'utilisation du véhicule
Informations liées aux véhicules	COD_MODELO	Code du modèle
	ANO_MODELO	Année de fabrication
	COD_TARIF	Code tarifaire
	COBERTURA	Couverture d'assurance
	IS_RCDMAT, IS_RCDC, IS_RCDMOR	Responsabilité civile (dommages matériels, corporels, moraux)
Variables cibles	INDENIZ	Montant de l'indemnité
	number_of_claims	Nombre de sinistres déclarés

Table 1: Présentation des variables utilisées pour l'analyse.

Pour la base des assurées la variable cible est `number_of_claims` qui représente le nombre de sinistres déclarés pour un contrat donné.Il s'agit d'une variable entière et non négative,essentielle pour analyser la fréquence des sinistres en fonction des caractéristiques des contrats et des assurés.
Pour la base des sinistres la variable cible est `INDENIZ` qui représente le montant de l'indemnité versé par l'assureur à la suite d'un sinistre déclaré. Il s'agit d'une variable continue et positive, essentielle pour analyser les classes influençant les coûts des sinistres.

2 Nettoyage et Explorations des données

2.1 Nettoyage des données

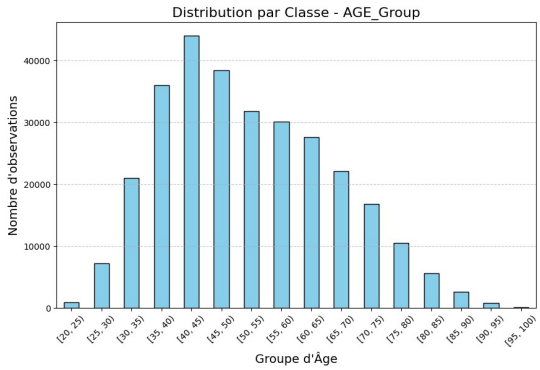
Le processus de néttoyage des données a été réalisé en plusieurs étapes successives afin de préparer la base `R_AUTO_2021A` pour la modélisation.
Tout d'abord, en raison de la taille importante de la base (31 940 158 lignes), celle-ci a été découpée en 64 fichiers de 500 000 lignes chacun, facilitant ainsi la manipulation des données. Ensuite, des colonnes non pertinentes, telles que `PRE_CASCO` (prime CASCO) et `COD_END` (type d'avenant), ont été supprimées pour réduire la complexité et la taille des fichiers.
Les données ont ensuite été standardisées : les formats de dates (`DATA_NASC` et `INICIO_VIG`) ont été convertis en format `datetime`, et les variables catégorielles comme `SEXO` (sexe) et `UTILIZACAO` (utilisation du véhicule) ont été encodées. Les valeurs aberrantes ont également été traitées, par exemple en supprimant les âges impossibles (> 124 ans) et en remplaçant les valeurs négatives dans les variables comme `IS_CASCO` par des valeurs manquantes (`NaN`).

Par la suite, de nouvelles variables ont été créées, notamment l'âge (`AGE`) calculé à partir de la date de naissance (`DATA_NASC`), et l'expérience du conducteur (`TEMPO_HAB`) convertie en années. Les fichiers nettoyés ont été fusionnés en une seule base, avec suppression des doublons et des incohérences, aboutissant à une base finale de 1 172 107 lignes.

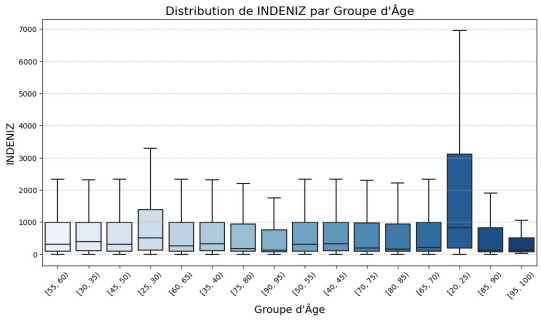
Enfin, une analyse exploratoire a permis de détecter d'éventuelles anomalies, telles que des âges élevés (> 100 ans) ou des montants aberrants dans les sommes assurées (`IS_CASCO`). Cette analyse a été réalisée à l'aide de visualisations, comme des boxplots et des histogrammes. À l'issue de ce processus, la base de données nettoyée contenait un ensemble réduit et pertinent de variables prêtes à être utilisées pour la modélisation.

Pour simplifier l'analyse en assurance automobile, une pratique courante est de manipuler que des variables qualitatives en créant des classes si nécéssaire. Nous avons alors regroupé l'âge des assurés en `classes d'âge` (par exemple : 25-30, 30-35, etc.), ce qui permet de faciliter l'interprétation et de rendre les modèles plus simples à appliquer. De manière similaire, nous avons créé des `classes d'années de modèle` des véhicules, en regroupant les années en intervalles cohérents. Ces regroupements permettent de simplifier les calculs tout en conservant la pertinence des variables pour la prédiction des sinistres. Cette approche est largement utilisée en assurance, où la priorité est donnée à la clarté et à l'interprétabilité des résultats.

2.2 Exploration des données



(a) Distribution de l'âge



(b) Indemnité par âge

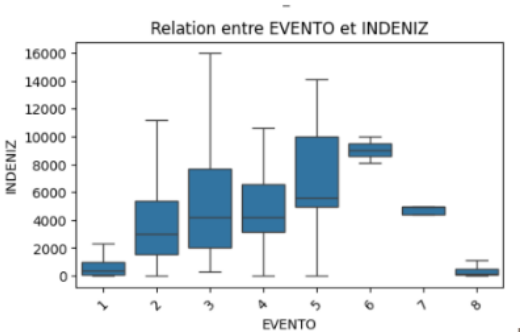
Figure 1: Exploration de l'âge

- **Distribution de l'âge (a) :** La répartition des assurés montre une concentration majeure dans les tranches d'âge $30-50$ ans, ce qui est cohérent avec une population active et propriétaire de véhicules. Les tranches plus jeunes (< 25 ans) et plus âgées (> 65 ans) sont significativement moins représentées, ce qui reflète probablement des comportements d'achat de véhicules ou des restrictions liées à l'usage.

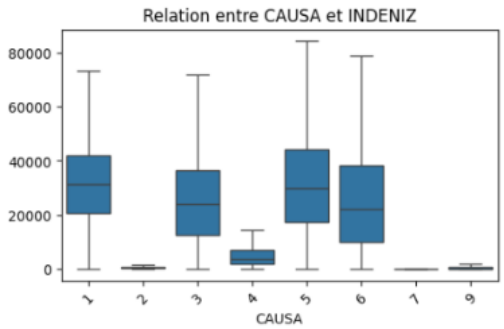
- **Indemnité par âge (b) :** Le groupe $20-25$ ans montre une plus grande dispersion des indemnités, avec des valeurs maximales bien plus élevées, probablement liées à des sinistres plus graves ou des comportements de conduite spécifiques à cette tranche d'âge.

  Les tranches d'âge plus élevées (> 70 ans) affichent des dispersions modérées mais globalement réduites, suggérant une fréquence moindre des sinistres ou des montants d'indemnisation généralement inférieurs.

Figure 2: Indemnité par Type d'accident



(a) Indemnité par type de sinistre



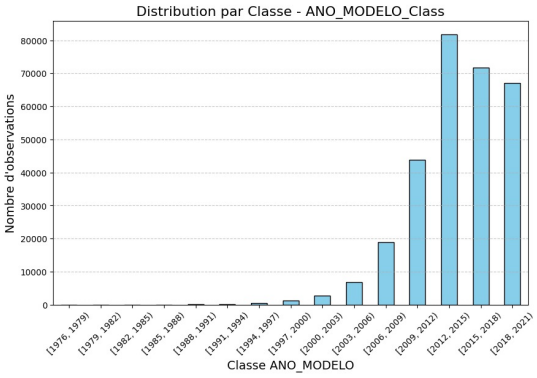
(b) Indemnité par cause de sinistre

- **(a) Indemnité par type de sinistre :** Les types de sinistres influencent directement les montants des indemnités. Par exemple, les sinistres liés aux "Accidents Personnels Passagers – Décès Accidentel" présentent des médianes et des dispersions plus élevées, ce qui reflète le coût important associé à ces types d'événements. En

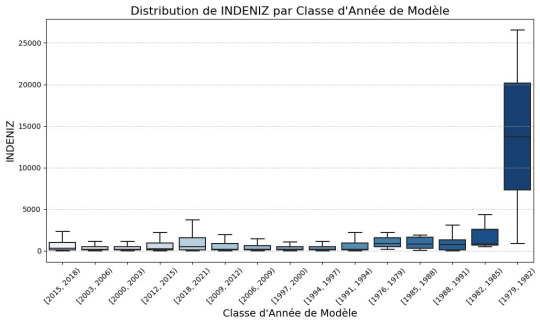
revanche, les sinistres de "Responsabilité Civile Facultative – Dommages Moraux" ont tendance à afficher des indemnités moins élevées.

- **(b) Indemnité par cause de sinistre :** Les causes des sinistres ont un impact marqué sur les montants des indemnités. Par exemple, les sinistres causés par des "Collisions avec indemnisation intégrale" ou des "Incendies" entraînent des montants significativement plus élevés en raison des dommages importants généralement associés. À l'opposé, les causes comme "Vol" ou "Assistance 24 heures" sont associées à des montants d'indemnisation plus modestes.

Figure 3: Indemnité par Année de modèle du véhicule



(a) Distribution de l'année du modèle du véhicule



(b) Indemnité par Année du modèle du véhicule

- **Distribution de l'année du modèle (a)** : Les véhicules récents (2009-2021) dominent largement, reflétant le renouvellement des automobiles. Les modèles plus anciens (avant 2000) sont marginalement représentés, en cohérence avec leur retrait progressif du marché actif.
- **Indemnité par année du modèle (b)** : Les modèles anciens (avant 1985) montrent des indemnités élevées et dispersées, probablement dues à des coûts de réparation plus élevés. Les modèles récents (après 2015) présentent des indemnisations plus homogènes, reflétant des avancées technologiques réduisant les coûts.

3 Modèles de machine learning

3.1 Evaluation des performances

Nous avons séparé notre base en deux jeux de données : l'un d'entraînement et l'autre de test. Les performances de nos modèles se mesurent sur le jeu de test. Le jeu d'entraînement a permis de mesurer la capacité du modèle à apprendre sur ce dernier. Pour une approche statistique comme le GLM, il faut pouvoir vérifier sa bonne adéquation à nos données. La performance d'un modèle dépend considérablement de ses hyperparamètres.

Cependant, la séparation unique en train-test présente certaines limites, notamment un risque de biais ou de variabilité élevée dans l'évaluation des performances en raison de

la dépendance au découpage choisi. Pour résoudre ce problème, nous avons également envisagé l'utilisation de la validation croisée en V-fold. Cette méthode consiste à diviser le jeu de données en V sous-ensembles de taille égale (ou folds). À chaque itération, un fold est utilisé comme ensemble de test, tandis que les $V - 1$ autres folds servent pour l'entraînement. Le processus est répété V fois, permettant ainsi à chaque sous-ensemble de servir une fois de test. Les résultats des V itérations sont ensuite moyennés pour obtenir une évaluation plus fiable et représentative des performances du modèle. Cette technique est particulièrement utile pour minimiser l'impact du découpage des données sur les résultats.

Il est important de noter qu'il n'est pas possible de prédire quelles seront les meilleures valeurs pour ces hyperparamètres, ce qui nécessite idéalement d'explorer toutes les valeurs possibles pour les déterminer. Faire cela manuellement serait une tâche laborieuse et coûteuse en termes de temps et de ressources, c'est pourquoi nous utilisons `GridSearchCV` pour automatiser ce processus de réglage des hyperparamètres. Comme son nom l'indique, on considère le problème d'optimisation comme un problème de recherche dans une grille. En pratique, on va simplement fixer pour chaque hyperparamètre un ensemble de valeurs qu'il peut prendre. Ensuite, pour chaque combinaison d'hyperparamètres, on va entraîner notre modèle et conserver les résultats de performances en mémoire. Il suffira donc de prendre les hyperparamètres pour lesquels les performances sont les meilleures.

En combinant cette approche avec la validation croisée en V-fold, nous garantissons une meilleure robustesse et généralisabilité des modèles entraînés. CV, cette approche assure à la fois robustesse et généralisabilité des modèles sur de nouvelles données.

3.2 Prédiction du coût des sinistres

3.2.1 GLM Gamma

Pour la modélisation de la sévérité, nous allons uniquement utiliser les sinistres ayant un coût strictement positif. La variable cible ici `INDENIZ`, qui est l'indemnisation versée pour le sinistre. Les variables explicatives utilisées dans cette analyse incluent les suivantes : `COD_MODELO` (modèle du véhicule), `ANO_MODELO` (année du modèle, variable numérique), `COD_TARIF` (catégorie tarifaire), `REGIAO` (région), `SEXO` (sexe du conducteur), `AGE` (âge du conducteur, variable numérique), `COBERTURA` (type de couverture), `EVENTO` (type de sinistre) et `CAUSA` (cause du sinistre). Afin de valider la pertinence des variables explicatives dans le cadre de la modélisation GLM, nous avons effectué une analyse de la variance (ANOVA).

Métrique	Valeur
Mean Absolute Error (MAE)	1405.32
Mean Squared Error (MSE)	1254821.47
Root Mean Squared Error (RMSE)	3542.81
R^2 Score	0.49
Déviance	2201738212019.12
AIC	1527894.23

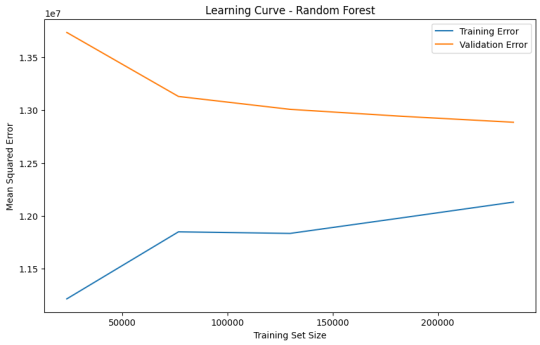
Table 2: Résultats des métriques du modèle.

3.2.2 RandomForest

Pour construire notre modèle, nous avons utilisé la fonction `RandomForestRegressor` de la bibliothèque scikit-learn. Le meilleur modèle a été identifié en optimisant les hyperparamètres tels que le nombre d'arbres (`n_estimators`) et la profondeur maximale (`max_depth`).

Métrique	Valeur
MSE moyen	12870510.33
MAE	1313.47
RMSE	3550.11
R^2	0.50
MSE	12603294.98
Déviance	1487138394752.78
AIC	973189.96

(a) Résultats des métriques.



(b) Learning curves RandomForest.

Figure 4: Résultats du modèle RandomForest.

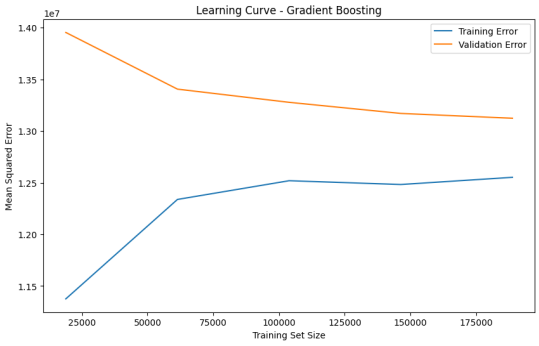
Le modèle RandomForest montre des performances correctes, mais perfectibles. Les métriques et les courbes d'apprentissage indiquent une certaine sur-adaptation aux données d'entraînement, ce qui peut limiter la capacité du modèle à généraliser sur de nouvelles données.

3.2.3 XGBoost

Pour la construction de notre modèle, nous avons utilisé la fonction XGBRegressor de la bibliothèque xgb. Le modèle optimal a été identifié grâce à l'optimisation de plusieurs hyperparamètres, notamment le nombre d'arbres (`n_estimators`), la profondeur maximale des arbres (`max_depth`) et le taux d'apprentissage (`learning_rate`). Les résultats obtenus sont présentés dans le tableau suivant :

Métrique	Valeur
MSE moyen	13131223.71
MAE	1343.90
MSE	12746650.18
RMSE	3570.25
R^2 Score	0.49
Déviance	1504053734111.20
AIC	973857.24

(a) Métriques du modèle.



(b) Courbes d'apprentissage XGBoost.

Figure 5: Résultats du modèle XGBoost.

3.3 Prédiction de la fréquence des sinistres

3.3.1 GLM Poisson

Pour la prédiction de la fréquence des sinistres, un modèle GLM a été construit en sélectionnant les 4 variables explicatives les plus significatives : le code tarifaire (`COD_TARIF`), la région géographique (`REGIAO`), l'âge du conducteur (`AGE_x`), et la garantie pour dommages matériels (`IS_RCDMAT`). Cette méthodologie a permis de réduire la complexité du modèle tout en maintenant de bonnes performances statistiques. Le modèle, basé sur un échantillon de 995,545 observations, présente une déviance observée de 883,487.10, inférieure à la valeur critique calculée (994,917.72), indiquant une adéquation satisfaisante entre les prédictions et les données réelles. La p-valeur associée de 1.0000 confirme que le modèle est statistiquement significatif. En conclusion, ce modèle simplifié conserve une bonne performance tout en expliquant efficacement la fréquence des sinistres.
En complément, une analyse par quantiles a été réalisée pour évaluer l'impact des valeurs extrêmes sur les métriques de performance. Les performances du modèle ont été comparées pour les quantiles 99%, 99.5%, et 99.9%. Cette démarche permet d'affiner l'évaluation du modèle en tenant compte des variations importantes dues aux sinistres rares mais coûteux.

Modèle	MAE	MSE	RMSE
GLM Poisson	1.0021	2.3483	1.5324

Table 3: Métriques d'évaluation pour la prédiction de la fréquence des sinistres avec le modèle GLM Poisson.

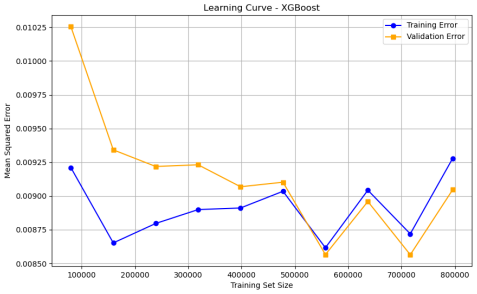
3.3.2 XGBoost

La méthodologie utilisée pour la modélisation s'appuie sur l'algorithme `XGBRegressor` de la bibliothèque `xgboost`. Une optimisation des hyperparamètres a été réalisée grâce à `GridSearchCV`, en testant divers paramètres tels que le nombre d'estimateurs (`n_estimators`),

la profondeur maximale des arbres (`max_depth`), le taux d'apprentissage (`learning_rate`), entre autres. Cette optimisation a inclus une validation croisée avec 3 plis (folds) et 243 combinaisons possibles d'hyperparamètres.
À l'issue de l'optimisation, les meilleurs hyperparamètres obtenus sont : un `colsample_bytree` de 1.0, un `learning_rate` de 0.1, une `max_depth` de 6, un `n_estimators` de 100 et un `subsample` de 0.8. Ces choix permettent d'assurer un bon équilibre entre performance et généralisation du modèle.

Métrique	Valeur
MAE	0.9742
MSE	2.1003
RMSE	1.4492

(a) Métriques du modèle.



(b) Courbes d'apprentissage XGBoost.

Figure 6: Résultats du modèle XGBoost.

3.4 Comparaisons des modèles

Pour le cout des sinistres :
Le RandomForest reste le modèle le plus performant globalement, grâce à sa précision et sa généralisation équilibrée. Le XGBoost offre des performances correctes, mais pourrait être amélioré par un réglage d'hyperparamètres. Le GLM Gamma, en revanche, est plus limité et montre des performances inférieures en raison de sa simplicité structurelle, ce qui le rend moins adapté à des données complexes avec des interactions non linéaires.

Métrique	RandomForest	XGBoost	GLM Gamma
MAE	1196.63	1313.41	1405.32
MSE	11095703.89	12070107.69	12548321.47
RMSE	3331.02	3474.21	3542.81
R^2	0.57	0.53	0.49
Déviance	1973770381919.85	2147103177011.09	2201738212019.12
AIC	1498177.44	1505664.13	1527894.23

Table 4: Comparaison des métriques entre RandomForest, XGBoost et GLM Gamma.

Pour la fréquence :
XGBoost est clairement supérieur au GLM pour prédire la fréquence des sinistres. Il bénéficie de sa capacité à capturer des relations non linéaires complexes entre les variables, tandis que GLM, malgré sa simplicité et son interprétabilité, est moins performant pour ce type de données.

Modèle	MAE	MSE	RMSE
GLM	1.0021	2.3483	1.5324
XGBoost	0.9742	2.1003	1.4492

Table 5: Comparaison des performances des modèles pour la prédiction de la fréquence des sinistres.

3.5 Estimation de la Prime pure

La prime pure a été estimée à l'aide d'une approche reposant sur le produit de la **fréquence** et du **coût moyen** des sinistres. Cette méthodologie, basée sur des modèles robustes, a permis une modélisation précise de la charge totale des sinistres.
Pour valider cette estimation, nous avons appliqué notre méthode sur un échantillon de **nouveaux assurés**. Les résultats obtenus se sont révélés **concluants**, démontrant la fiabilité et la pertinence de notre approche. Les estimations obtenues sont cohérentes avec les données réelles, confirmant que cette méthode est bien adaptée pour prédire les charges potentielles liées aux sinistres (en Réal brésilien annuellement) :

Age	Region	Type véhicule	Coût	Fréquence	Prime Pure
20-25	Brasília	Véhicules tourismes	4070.25	1.93	7855.58
30-35	Toledo-cascavel	Véhicules tourismes	1370.15	2.75	3767.91
35-40	BA - Bahia	Bicyclette, motos	161.64	1.67	269.93
40-45	Ribeirão Preto	Pick up lourd	4156.01	0.30	1246.80
55-60	Met. Porto Alegre	Pick up lourd	1781.81	0.20	356.36

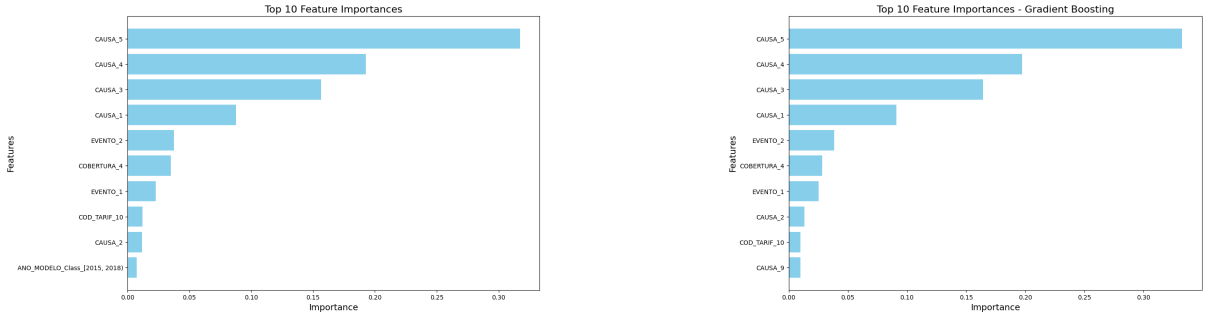
Table 6: Tableau des prédictions : Coût, fréquence et prime pure.

Nous remarquons alors que pour la tranche d'âge 20-25, les primes prédites sont très élevées. Cela est conforme à notre portefeuille comme en témoigne le boxplot de la figure 1, à la page 6.
Par ailleurs, étant des nouveaux assurés, ils n'ont pas d'historique de sinistre réduisant alors énormément la prime pure prédite.

Finalement, notre modèle reste large, ainsi, avec le peu d'information que nous avons, nous pouvons toujours fournir une estimation de la prime pure conformément à notre portefeuille.

4 Interpretations

4.1 Importance des variables



(a) RandomForest (b) XGBoost

Figure 7: Importance des variables

Les résultats montrent que les `causes des sinistres (CAUSA)` et les `types de sinistres (EVENTO)` jouent un rôle clé dans l'explication des montants d'indemnisation (`INDENIZ`). Parmi les causes, `CAUSA_5` (collisions avec indemnisation intégrale) est la plus influente, soulignant la gravité de ces sinistres qui impliquent des dommages matériels importants nécessitant une couverture complète. Ensuite, `CAUSA_4` (collisions partielles) et `CAUSA_3` (rapts) démontrent que les collisions fréquentes, même partielles, et les incidents graves comme les rapts entraînent également des coûts élevés. Du côté des types de sinistres, `EVENTO_2` (Responsabilité Civile pour Dommages Matériels) est particulièrement significatif, reflétant les coûts liés à la prise en charge des dommages causés à des tiers.

4.2 Shap

SHAP est une méthode qui permet de comprendre comment un modèle utilise les variables pour faire ses prédictions. En attribuant une importance à chaque classe de chaque variable pour chaque prédiction, SHAP aide à identifier quelles variables influencent le plus les résultats. Rappelons que dans notre étude de tarification, nous n'avons plus de variable explicative quantitative. Etant une pratique courante chez les assureurs pour faciliter les études de tarification. Nous présentons alors seulement un échantillon en soulignant les classes prédominants sur notre prédiction:

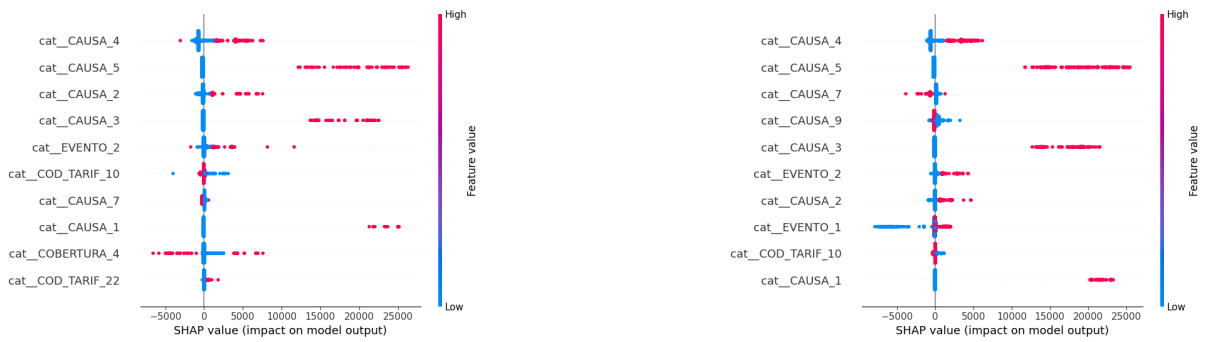


Figure 8: RandomForest et XGBoost

Les graphiques SHAP permettent d'analyser en détail l'impact individuel de chaque classe de chaque variable sur les prédictions des modèles `Random Forest` et `XGBoost`, offrant une compréhension plus fine des relations entre les caractéristiques et les montants d'indemnisation (`INDENIZ`). Les variables `CAUSA_4` (collisions partielles) et `CAUSA_5` (collisions avec indemnisation intégrale) se distinguent par leur influence majeure, où les valeurs élevées de ces variables (en rouge) augmentent considérablement les montants prévus, soulignant leur lien avec des sinistres graves et coûteux. `CAUSA_3` (rapts), bien que légèrement moins influente dans le modèle `Random Forest`, montre un impact significatif dans le modèle `XGBoost`, mettant en évidence le coût élevé des sinistres graves comme les enlèvements. Les types de sinistres liés à la responsabilité civile, en particulier `EVENTO_2` (dommages matériels), jouent également un rôle clé, leurs valeurs élevées augmentant notablement les prédictions, ce qui reflète les coûts élevés liés à la couverture de ces dommages. Les catégories tarifaires spécifiques, telles que `COD_TARIF_10`(véhicules de touristes nationaux), et les couvertures étendues, comme `COBERTURA_4`(couverture intégrale), contribuent également à expliquer les montants d'indemnisation, bien que leur impact soit plus modéré. La cohérence entre les deux modèles dans l'identification des variables les plus influentes renforce la robustesse des résultats, montrant que les sinistres graves et les caractéristiques associées dominent l'explication des coûts pour les assureurs.

Maintenant, pour la fréquence, nous regroupons les classes de chaque variable pour avoir l'impact global de la variable sur la prédiction. Le risque en faisant cela est de prendre les informations de chaque classe de chaque variable et d'avoir des compensations entre les classes d'une variable. Faisons le quand même pour illustrer l'étude :

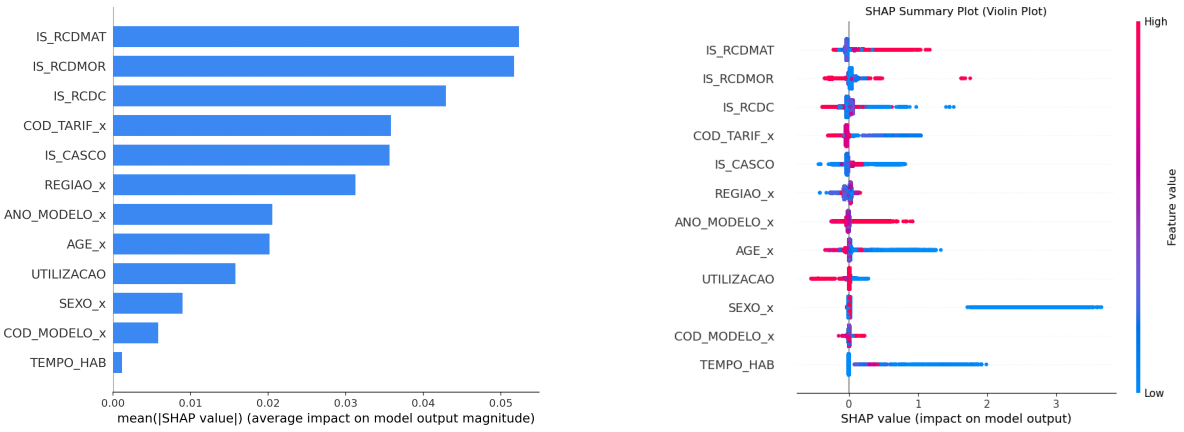


Figure 9: XGBoost de la fréquence

- **Importance moyenne des variables :**
  - Les variables `IS_RCDMAT`, `IS_RCDMOR`, et `IS_RCDC` dominent en termes d'impact sur les prédictions, ce qui indique que les garanties de responsabilité civile (dommages matériels, moraux et corporels) sont les principaux classes influençant la fréquence des sinistres.
  - D'autres variables comme `COD_TARIF_X` (code tarifaire), `IS_CASCO` (garantie casco), et `REGIAO_X` (région géographique) contribuent également de manière significative, mais dans une moindre mesure.
- **Diagramme de dispersion SHAP :**

- Le diagramme de dispersion révèle non seulement l'importance des variables, mais également leur influence positive ou négative sur les prédictions.
- Par exemple, pour `IS_RCDMAT`, des valeurs élevées augmentent la fréquence prédite, tandis que des valeurs plus faibles ont un effet inverse.
- La variable `TEMPO_HAB` (durée d'habitation) semble avoir un impact plus modéré mais reste informative.

Les garanties liées aux sinistres (responsabilité civile et casco) et des classes comme le code tarifaire et la région sont déterminants dans la prédiction de la fréquence des sinistres. Ces résultats confirment la pertinence des variables sélectionnées et la capacité du modèle `XGBoost` à bien interpréter les données pour ce problème.

4.3 Dépendance Partiel

Il s'agit d'une méthode globale, elle vise à montrer l'effet marginal d'une ou de plusieurs variables sur les prédictions réalisées par le modèle. Cette méthode consiste à regrouper en deux groupes les variables explicatives, le premier groupe est celui qui contient les variables explicatives (pas plus de 2 généralement) pour lesquelles on veut connaitre l'effet sur la prédiction, et le second groupe contient les variables explicatives restantes, pour ensuite calculer une fonction appelée fonction de dépendance partielle.
Dans notre situation, avec seulement des variables explicatives qualitatives, nous présentons un exemple avec 3 classes de trois de nos variables qualitatives ayant le plus d'impact sur la prédiction de coût.

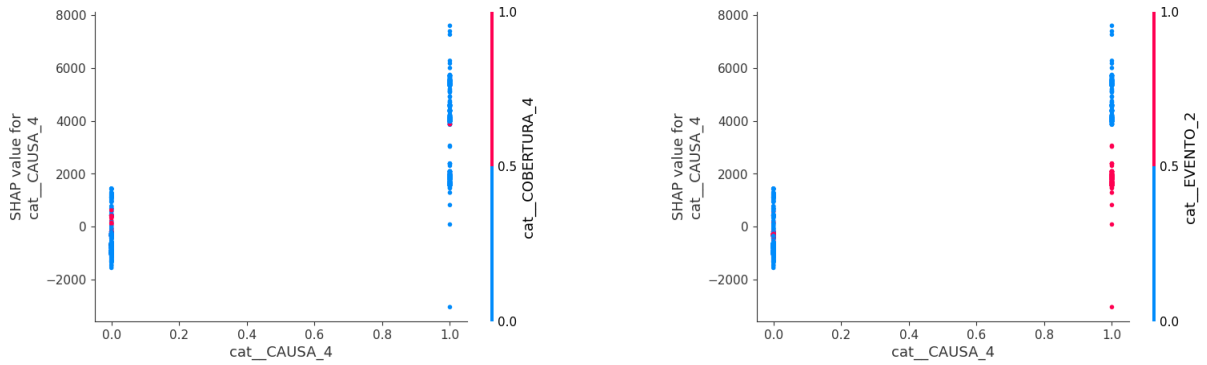


Figure 10: Dépendance Partiel RF (coût)

Pour la première image, on observe que lorsque `CAUSA 4` (collision partielle) est activée en même temps que `COBERTURA 4` (indemnisation intégrale, collision et vol), la prédiction d'`INDENIZ` (indemnisation) augmente de manière significative, comme le montrent les points rouges situés dans les valeurs élevées des SHAP values. Cette combinaison reflète des situations où l'assuré bénéficie d'une couverture maximale pour des collisions partielles, ce qui entraîne des coûts prévus élevés pour l'assureur.
En revanche, lorsque `COBERTURA 4` n'est pas activée (points bleus), l'effet de `CAUSA 4` sur la prédiction est moins important, voire proche de zéro. Cela suggère que la couverture joue un rôle essentiel pour expliquer les coûts élevés liés aux collisions partielles.

Pour la deuxième image, lorsque `CAUSA 4` (collision partielle) est combinée avec `EVENTO`

2 (responsabilité civile facultative – dommages matériels), les prédictions d'INDENIZ augmentent de manière notable, comme indiqué par les points rouges. Cela montre que les collisions partielles, lorsqu'elles impliquent des dommages matériels pris en charge par la responsabilité civile, sont des situations coûteuses pour l'assureur.
En revanche, pour les points bleus (cas où EVENTO 2 n'est pas activé), l'effet de CAUSA 4 sur la prédiction est beaucoup plus faible. Cela illustre que, dans l'absence d'un événement de responsabilité civile dommages matériels, les collisions partielles n'ont pas un impact significatif sur les prédictions d'indemnisation.

Maintenant pour la fréquence, nous donnons un exemple en regroupant les classes de la variable AGE et le code du tarif. Le risque en faisant cela est de perdre les informations de chaque classe des variables associées.

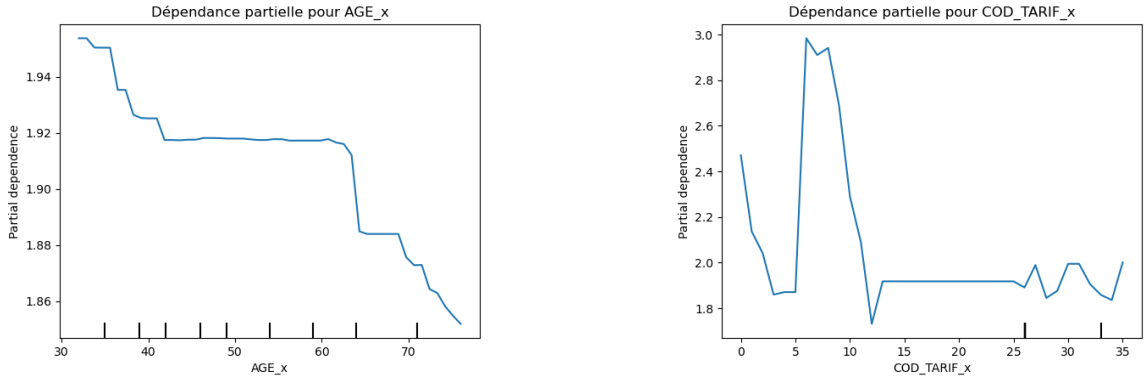


Figure 11: Dépendance Partiel XGBoost (fréquence)

Ces résultats mettent en évidence les effets spécifiques des variables AGE et COD_TARIF sur la fréquence prédite des sinistres. Tandis que l'âge présente une relation plus linéaire et intuitive, les codes tarifaires introduisent une complexité qui reflète probablement des différences structurelles ou catégoriques dans les données d'assurance. Ces observations confirment la pertinence de ces variables dans le modèle.

5 Conclusion

Le projet menée montre que les méthodes de machine learning offrent des performances comparables à celles des modèles linéaires. Dans la plupart des structures, la détermination des tarifs automobile repose aujourd'hui exclusivement sur des modèles linéaires généralisés GLM qu'il est intéressant de challenger par différentes approches, pour déterminer au cas par cas la plus pertinente. Il n'est reste pas moins que le GLM est mieux compris par de nombreux opérateurs et plus facile pour certains à insérer dans leurs systèmes de gestion et dans leurs OAV.