Institut National de Statistique et d'Économie Appliquée

Haut-Commissariat au Plan

Élément :

Régression en grande dimension

Détection de la fraude à l'assurance

Approche par des modèles de machine learning supervisés

Réalisé par :

- KHABIR Safia
- BOUATMANE Kaouthar
- EL MENOUAR Adam
- CAIDI Yassine

Encadré par :
Mr. JANATI Hicham

Année universitaire : 2024 – 2025

1 Introduction

La fraude à l'assurance représente un problème économique majeur pour les compagnies d'assurance. Elle engendre des pertes financières importantes, une augmentation des coûts opérationnels et une dégradation de la relation de confiance avec les assurés. Face à la croissance du volume des sinistres et à la complexité des mécanismes de fraude, les méthodes traditionnelles de détection, basées sur des règles fixes ou des analyses manuelles, montrent leurs limites.

La détection automatique de la fraude constitue un défi important en raison du fort déséquilibre entre les sinistres frauduleux et non frauduleux, de la diversité des comportements observés et du risque élevé de faux positifs. Un système de détection inefficace peut conduire soit à laisser passer des fraudes coûteuses, soit à pénaliser des assurés légitimes, ce qui souligne la nécessité de solutions performantes et interprétables.

Problématique

Dans ce contexte, la problématique principale de ce projet est la suivante : *comment concevoir un modèle de machine learning capable de détecter efficacement les sinistres frauduleux dans un environnement fortement déséquilibré, tout en assurant un compromis pertinent entre performance, robustesse et interprétabilité des décisions ?*

Objectifs du projet

L'objectif général de ce travail est de développer et d'évaluer une approche de détection de la fraude basée sur des techniques de machine learning. Plus précisément, les objectifs sont les suivants :

- analyser les caractéristiques des données de sinistres et identifier les variables potentiellement discriminantes ;
- mettre en place un prétraitement adapté, notamment pour le traitement du déséquilibre des classes ;
- comparer plusieurs modèles de classification à l'aide de métriques adaptées ;
- sélectionner le modèle le plus performant à l'aide d'une validation croisée ;

- évaluer le modèle retenu sur des données réelles et analyser l'impact du seuil de décision dans un contexte métier ;
- interpréter les prédictions du modèle afin de mieux comprendre les facteurs influençant la détection de la fraude.

2 Présentation et description des données

Les données utilisées dans ce projet proviennent de la base publique *Insurance Claims Fraud Dataset*, disponible sur la plateforme Mendeley Datared[1]. Ce jeu de données contient 1000 observations correspondant à des déclarations de sinistres automobiles, décrites à l'aide de variables hétérogènes d'ordre démographique, contractuel, financier et événementiel.

La variable cible du modèle est `fraud_reported`, une variable binaire indiquant si le sinistre a été déclaré comme frauduleux (`Y`) ou non (`N`). Les variables explicatives peuvent être regroupées en plusieurs catégories principales :

- **Variables socio-démographiques** : âge, sexe de l'assuré, relation avec le souscripteur, ancienneté du client (`months_as_customer`).
- **Variables contractuelles** : montant de la prime annuelle, franchise (`policy_deductible`), limite de garantie (`umbrella_limit`), État de souscription de la police.
- **Variables liées au sinistre** : type et gravité de l'incident, heure de survenance, nombre de véhicules impliqués, présence de témoins, disponibilité du rapport de police.
- **Variables financières** : montants des indemnisations (corporelles, matérielles et véhicule), gains et pertes en capital.
- **Variables liées au véhicule** : marque du véhicule, type de collision et dommages matériels.

La diversité et la forte dimension de ces variables rendent ce jeu de données particulièrement adapté à l'application de méthodes de classification supervisée en grande dimension pour la détection de la fraude à l'assurance.

[1] magenta`https://data.mendeley.com/datasets/992mh7dk9y/2`

3 Nettoyage des données

Avant la phase de modélisation, un nettoyage de base des données a été réalisé afin d'améliorer la qualité du jeu de données et de supprimer les informations non pertinentes ou redondantes.

Dans un premier temps, une colonne technique sans signification métier a été supprimée. Ce type de variable, généralement issu de problèmes d'exportation ou de formatage, n'apporte aucune information utile à la modélisation et peut introduire du bruit.

Ensuite, certaines variables catégorielles comportant des valeurs manquantes ont été traitées. Les valeurs absentes ont été remplacées par la modalité *Unknown*, permettant de conserver l'information liée à l'absence de déclaration sans supprimer d'observations. Cette stratégie est particulièrement adaptée aux variables telles que le type de collision, la présence de dommages matériels ou la disponibilité d'un rapport de police.

Par ailleurs, plusieurs variables ont été supprimées car elles étaient peu informatives ou constituaient de simples identifiants. Il s'agit notamment du numéro de police, de la localisation textuelle du sinistre, du code postal de l'assuré et de certaines variables strictement techniques. Ces variables ne contribuent pas directement à la détection de la fraude et peuvent biaiser l'apprentissage du modèle.

À l'issue de ce nettoyage, le jeu de données contient 1000 observations et 34 variables. Une vérification finale a confirmé l'absence totale de valeurs manquantes, garantissant ainsi un jeu de données propre et exploitable pour les étapes suivantes de prétraitement et de modélisation.

4 Analyse exploratoire des données

L'analyse exploratoire des données (EDA) a été réalisée afin de mieux comprendre la structure du jeu de données, d'identifier les caractéristiques des sinistres frauduleux et de guider les choix de prétraitement et de modélisation.

4.1 Distribution de la variable cible

La distribution de la variable cible `fraud_reported` met en évidence un fort déséquilibre entre les classes. Les sinistres non frauduleux représentent une large majorité des observations, tandis que les sinistres frauduleux constituent une minorité du jeu de données. Ce déséquilibre justifie l'utilisation de métriques adaptées et de techniques spécifiques pour le traitement des classes minoritaires.

4.2 Analyse des variables numériques

L'étude des distributions des variables numériques montre des comportements hétérogènes. La variable `months_as_customer` présente une distribution étalée, indiquant une grande variabilité dans la durée de relation entre les assurés et la compagnie. La variable `age` suit une distribution proche d'une loi normale, avec une concentration des assurés entre 30 et 45 ans.

Les variables financières, telles que `policy_annual_premium`, `capital_gains` et `capital_loss`, présentent des distributions fortement asymétriques, avec la présence de valeurs extrêmes. La variable `umbrella_limit` montre une concentration importante de valeurs nulles, associée à quelques valeurs très élevées, ce qui justifie un traitement particulier lors du prétraitement.

L'analyse de la matrice de corrélation révèle une faible corrélation entre la majorité des variables numériques, à l'exception des composantes du montant du sinistre (`injury_claim`, `property_claim` et `vehicle_claim`), qui sont fortement corrélées entre elles. Cette observation suggère une redondance partielle de l'information financière.

4.3 Analyse des variables catégorielles

L'analyse bivariée entre les variables catégorielles et la variable cible met en évidence des différences notables entre sinistres frauduleux et non frauduleux. Les sinistres de type *Major Damage* et *Total Loss* présentent une proportion plus élevée de fraudes comparativement aux sinistres de faible gravité.

De même, certains types d'incidents, notamment les collisions impliquant plusieurs véhicules, sont plus fréquemment associés à des cas de fraude. Le type de collision (*Rear Collision*, *Side Collision*) montre également des variations dans la répartition des fraudes.

La présence ou l'absence de dommages matériels déclarés, ainsi que la disponibilité d'un rapport de police, semblent également jouer un rôle dans la distinction entre sinistres frauduleux et non frauduleux. En revanche, des variables telles que la marque du véhicule ou le sexe de l'assuré ne présentent pas de différences marquées entre les deux classes.

4.4 Traitement des valeurs extrêmes

L'analyse exploratoire a mis en évidence une forte asymétrie de la variable *umbrella_ limit*. Les statistiques descriptives montrent que plus de 75% des observations sont nulles, tandis que certaines valeurs atteignent des montants très élevés, jusqu'à 10 millions. Cette distribution très déséquilibrée est caractéristique des garanties complémentaires rarement souscrites, mais pouvant atteindre des montants importants.

Les boxplots réalisés avant traitement confirment la présence de valeurs extrêmes qui écrasent la distribution et rendent difficile l'analyse de la partie centrale. Ces valeurs aberrantes sont susceptibles d'influencer négativement l'apprentissage des modèles, en particulier les modèles sensibles à l'échelle des variables.

Afin de limiter l'impact de ces observations extrêmes sans les supprimer, une stratégie de capping au 99^{e} centile a été appliquée. Toutes les valeurs supérieures à ce seuil ont été ramenées à la valeur du 99^{e} centile. Cette approche permet de conserver l'information contenue dans les sinistres rares tout en stabilisant la distribution de la variable.

Les boxplots après capping montrent une distribution plus compacte et mieux adaptée à la phase de modélisation, tout en respectant la logique métier du domaine assurantiel.

5 Préparation des données pour la modélisation

La variable cible *fraud_ reported* a été encodée sous forme binaire, où la modalité Y correspond aux sinistres frauduleux et la modalité N aux sinistres non frauduleux. Les variables explicatives ont ensuite été séparées de la cible afin de constituer la matrice des caractéristiques.

Les variables de type date (*policy_ bind_ date*, *incident_ date* et *date_ of_ birth*) ont été supprimées, leur exploitation nécessitant un traitement spécifique de type ingénierie de caractéristiques qui dépasse le cadre de ce projet.

Les variables catégorielles ont été transformées par encodage *one-hot*, permettant leur conversion en variables numériques binaires. Après encodage, le jeu de données comprend 142 variables explicatives.

Les données ont ensuite été séparées en un ensemble d'apprentissage (70%) et un ensemble de test (30%), en conservant la proportion initiale de sinistres frauduleux grâce à une séparation stratifiée. L'analyse de la répartition de la cible dans l'échantillon d'apprentissage met en évidence un fort déséquilibre entre les classes, avec environ 25% de sinistres frauduleux.

Afin de corriger ce déséquilibre, une standardisation des variables numériques a été réalisée à l'aide d'un *StandardScaler*, suivie de l'application de la méthode SMOTE sur le seul ensemble d'apprentissage. Cette technique de sur-échantillonnage synthétique permet d'obtenir un jeu d'apprentissage équilibré, tout en évitant toute fuite d'information vers l'ensemble de test.

5.1 Définition des modèles et stratégie d'évaluation

Quatre modèles de classification ont été retenus afin de comparer différentes approches de détection de la fraude. Une régression logistique a été utilisée

comme modèle linéaire de référence. Un classifieur SVM à noyau radial a été sélectionné pour sa capacité à modéliser des frontières de décision non linéaires. Un Random Forest a été employé pour exploiter les interactions complexes entre les variables. Enfin, un classifieur par vote (*Voting Classifier*) a été construit afin de combiner les prédictions des modèles précédents.

Pour chaque algorithme, le paramètre *class_ weight='balanced'* a été activé afin de tenir compte du déséquilibre initial des classes. Le SVM a été configuré pour produire des probabilités, permettant ainsi le calcul de la courbe ROC et l'analyse de seuil.

L'évaluation des modèles a été réalisée à l'aide d'une fonction unique, garantissant une comparaison équitable. Chaque modèle a été entraîné sur le jeu d'apprentissage rééquilibré par SMOTE, puis évalué sur le jeu de test non modifié. Les métriques retenues incluent l'accuracy, la précision, le rappel, le F1-score et l'aire sous la courbe ROC (ROC-AUC). Un rapport de classification détaillé est également produit afin d'analyser les performances par classe.

5.2 Entraînement des modèles et comparaison des performances

Table 1: Comparaison des performances des modèles sur le jeu de test

Modèle	**Accuracy**	**Précision**	**Rappel**	**F1-score**	**ROC-AUC**
Régression logistique	0.800	0.621	0.486	**0.545**	**0.826**
Voting Classifier	0.790	**0.634**	0.351	0.452	0.816
Random Forest	0.770	0.568	0.284	0.378	0.778
SVM	0.757	0.526	0.135	0.215	0.760

Le tableau **??** présente les performances des différents modèles évalués sur le jeu de test. La régression logistique se distingue par le meilleur F1-score et la meilleure aire sous la courbe ROC, traduisant un bon compromis entre la précision et le rappel pour la détection des fraudes.

Le classifieur par vote obtient la meilleure précision, mais son rappel reste inférieur, ce qui limite sa capacité à détecter l'ensemble des sinistres frauduleux. Le Random Forest et le SVM affichent des performances globales plus faibles, en particulier en termes de rappel, indiquant une sous-détection significative des fraudes lorsque le seuil de décision par défaut est utilisé.

Ces résultats constituent une première évaluation comparative, mais restent dépendants d'un découpage unique des données. Une validation croisée est donc nécessaire afin d'évaluer la robustesse et la stabilité des performances des modèles.

5.2.1 Matrices de confusion

Les matrices de confusion montrent que la régression logistique et le classifieur par vote offrent un meilleur équilibre entre faux positifs et faux négatifs. Le SVM privilégie fortement la classe majoritaire, entraînant une sous-détection importante des sinistres frauduleux. Le Random Forest améliore la détection par rapport au SVM, mais reste limité par un rappel insuffisant avec le seuil par défaut.

5.2.2 Courbes ROC

Les courbes ROC confirment la capacité de discrimination des modèles, la régression logistique présentant la meilleure aire sous la courbe. Le Voting Classifier et le Random Forest affichent des performances proches, tandis que le SVM reste en retrait. Ces résultats justifient une analyse du seuil de décision afin d'adapter le modèle aux contraintes métier.

6 Validation croisée multi-métriques

Afin d'évaluer la robustesse et la stabilité des modèles, une validation croisée stratifiée à cinq plis est appliquée sur les données d'entraînement rééquilibrées par SMOTE. Les performances sont mesurées à l'aide de plusieurs métriques complémentaires, à savoir l'accuracy, la précision, le rappel, le F1-score et l'aire sous la courbe ROC. Cette approche permet de limiter la dépendance à un découpage unique des données et d'obtenir une estimation plus fiable des performances réelles des modèles.

Table 2: Performances moyennes des modèles après validation croisée (5-fold)

Modèle	**Accuracy**	**Précision**	**Rappel**	**F1-score**	**ROC-AUC**
Régression logistique	0.888	0.910	0.863	0.885	0.959
SVM	0.917	**0.963**	0.869	0.913	0.978
Random Forest	**0.929**	0.951	0.905	**0.927**	**0.980**
Voting Classifier	0.922	0.936	**0.907**	0.921	0.979

Les performances moyennes issues de la validation croisée montrent que le Random Forest obtient les meilleurs résultats globaux, notamment en termes de F1-score et de capacité de discrimination, et est donc retenu comme modèle final.

6.1 Évaluation finale sur le jeu de test

Le modèle Random Forest sélectionné à l'issue de la validation croisée a été réentraîné sur l'ensemble des données d'apprentissage rééquilibrées par

SMOTE, puis évalué sur le jeu de test conservé dans sa distribution réelle. Les résultats obtenus avec le seuil de classification par défaut (0.5) montrent une accuracy de 0.77 et un ROC-AUC de 0.778. Toutefois, le rappel de la classe frauduleuse reste faible, indiquant une sous-détection des sinistres frauduleux lorsque le seuil standard est utilisé.

6.2 Analyse de seuil

Une analyse de seuil a été menée afin d'adapter le comportement du modèle aux contraintes métier de la détection de fraude. Les résultats montrent qu'une diminution du seuil de décision permet d'augmenter significativement le rappel, au prix d'une baisse de la précision. Un seuil de 0.4 constitue un compromis pertinent, tandis qu'un seuil de 0.3 peut être privilégié dans un contexte où la priorité est de minimiser les fraudes non détectées.

6.3 Interprétation du modèle par SHAP

Afin d'interpréter les décisions du modèle Random Forest, la méthode SHAP a été utilisée pour quantifier l'impact de chaque variable sur la prédiction de fraude. L'analyse globale montre que la sévérité du sinistre, certains types de collisions, la présence de dommages matériels non renseignés ainsi que plusieurs caractéristiques de l'assuré figurent parmi les variables les plus influentes.

Le diagramme en barres met en évidence l'importance moyenne des variables, tandis que le graphique de dispersion (beeswarm) illustre la direction et l'amplitude de leur influence. Ces résultats confirment que le modèle s'appuie sur des facteurs cohérents d'un point de vue métier, renforçant ainsi la confiance dans ses prédictions.

References

[1] Dal Pozzolo, A., Bontempi, G., Snoeck, M., *Adversarial Drift Detection in Fraud Analytics*, IEEE Transactions on Neural Networks and Learning Systems, 2015.

[2] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd Edition, 2009.

[3] Breiman, L., *Random Forests*, Machine Learning, vol. 45, pp. 5–32, 2001.

[4] Cortes, C., Vapnik, V., *Support-Vector Networks*, Machine Learning, vol. 20, pp. 273–297, 1995.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[6] Lundberg, S. M., Lee, S.-I., *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems (NeurIPS), 2017.

[7] Fawcett, T., *An Introduction to ROC Analysis*, Pattern Recognition Letters, vol. 27, pp. 861–874, 2006.

[8] Mendeley Data, *Insurance Claims Fraud Dataset*, Version 2, 2020. Disponible en ligne : magenta`https://data.mendeley.com/datasets/992mh7dk9y/2`

[9] Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[10] He, H., Garcia, E. A., *Learning from Imbalanced Data*, IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, 2009.

Annexes

A Analyse exploratoire des données

A.1 Distribution de la variable cible

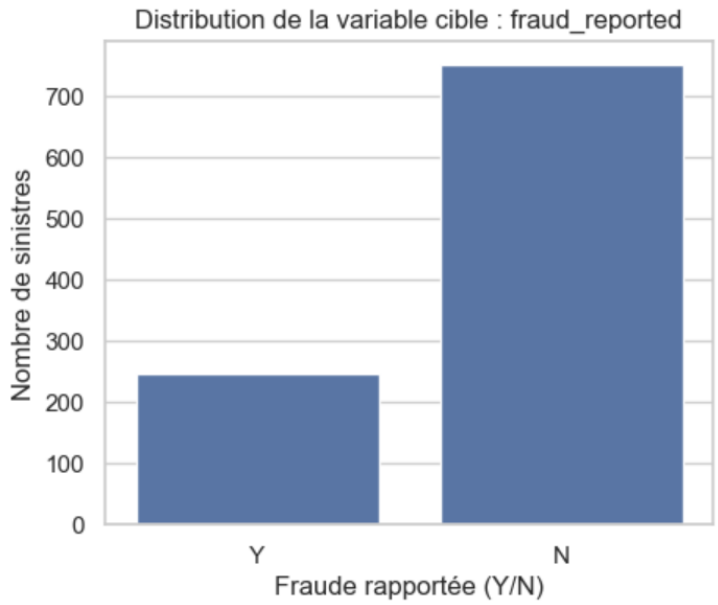

Figure 1: Distribution de la variable cible *fraud_ reported*

A.2 Distributions des variables numériques clés

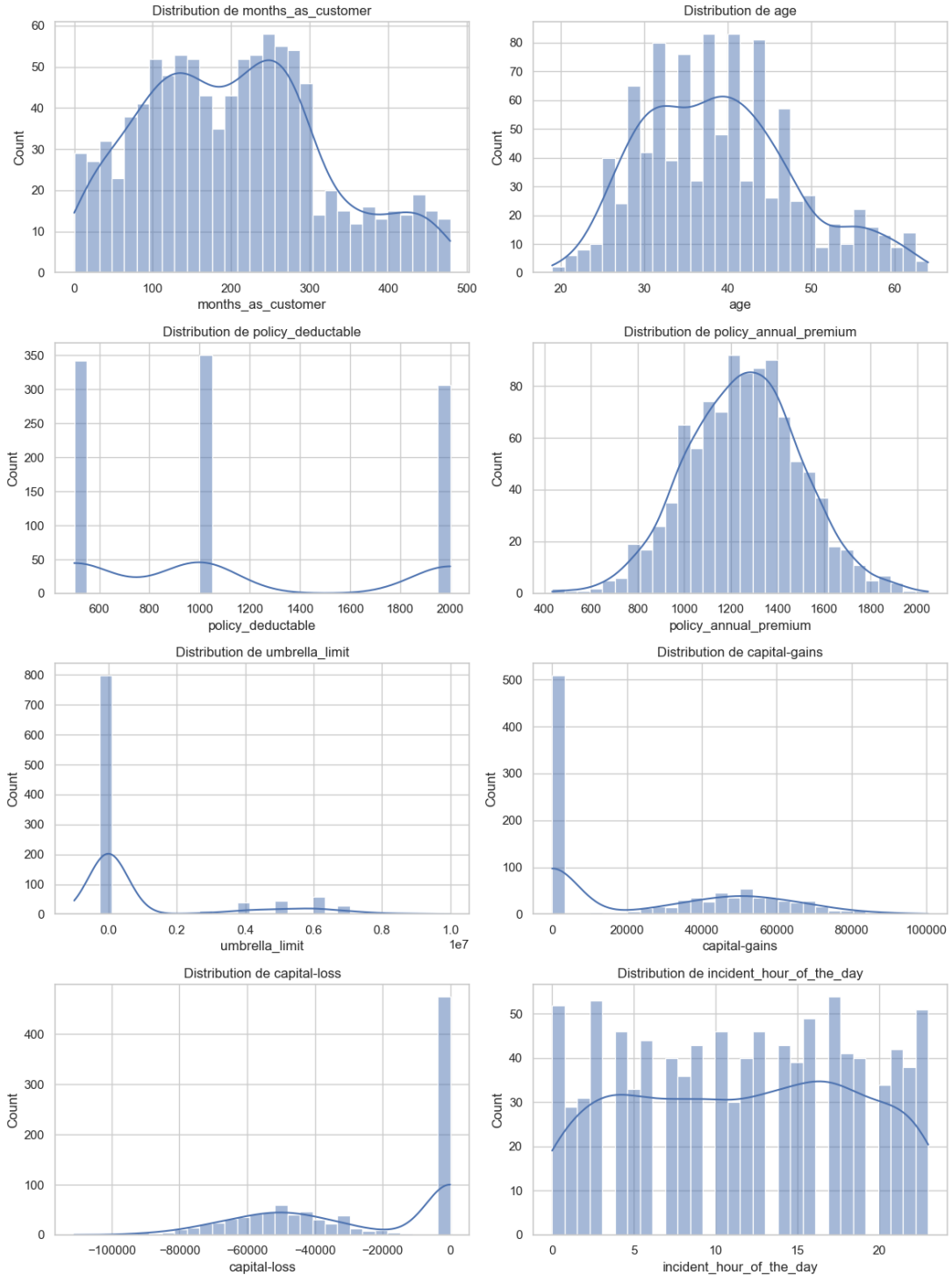

Figure 2: Distributions des principales variables numériques

A.3 Matrice de corrélation

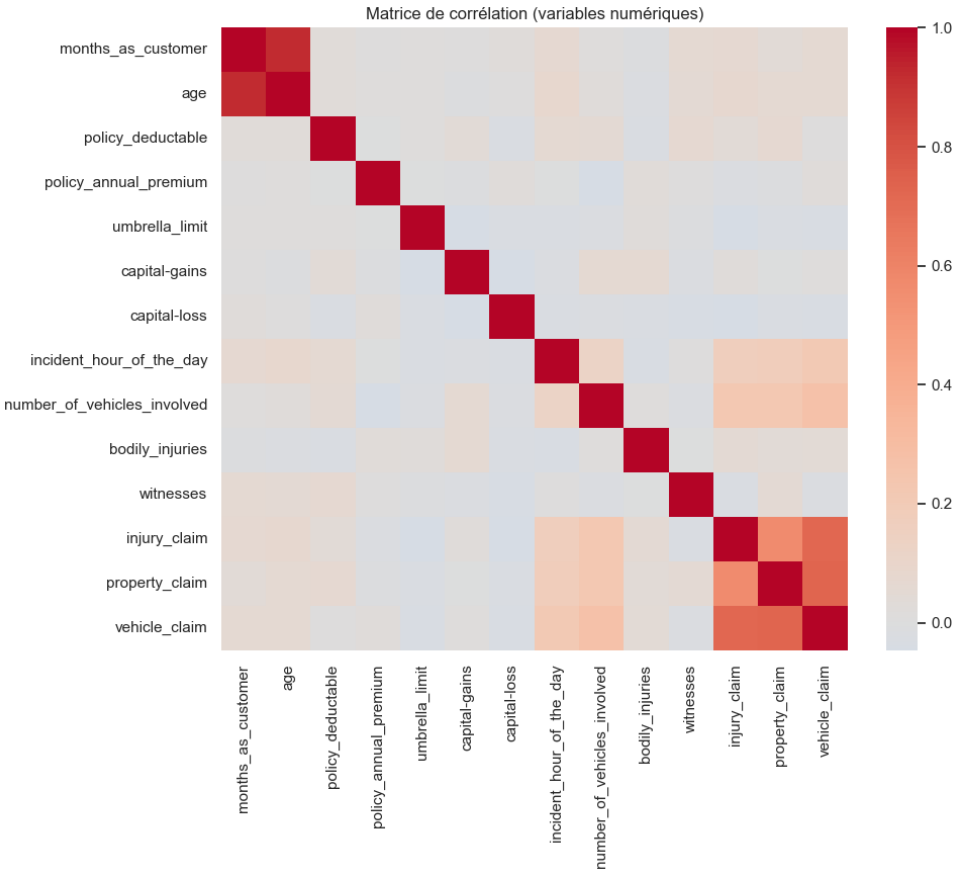

Figure 3: Matrice de corrélation des variables numériques

B Prétraitement et traitement des valeurs extrêmes

B.1 Traitement des outliers de *umbrella_limit*

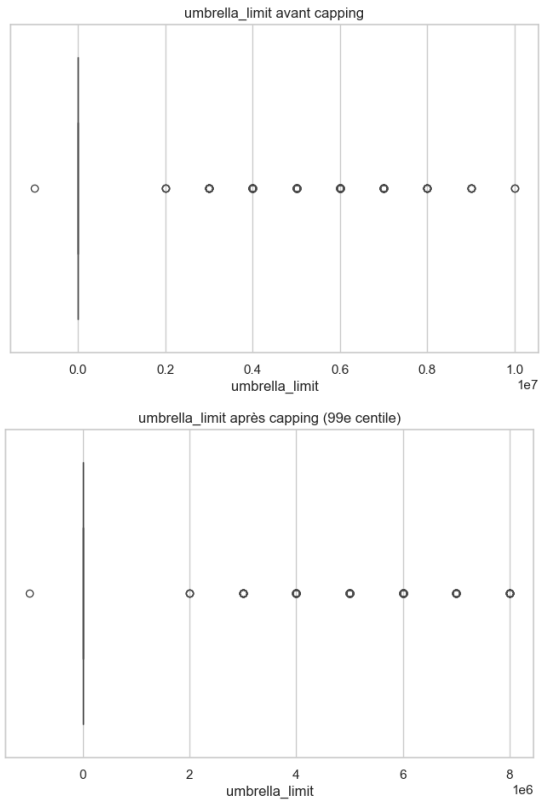

Figure 4: Traitement des valeurs extrêmes de *umbrella_limit* avant et après capping

C Évaluation des modèles

C.1 Matrices de confusion

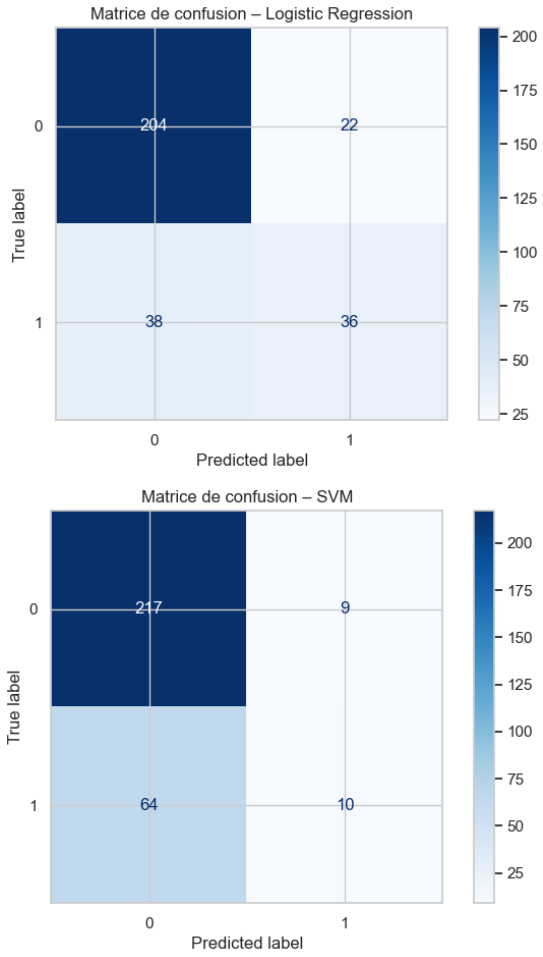

Figure 5: Matrices de confusion : Régression Logistique et SVM

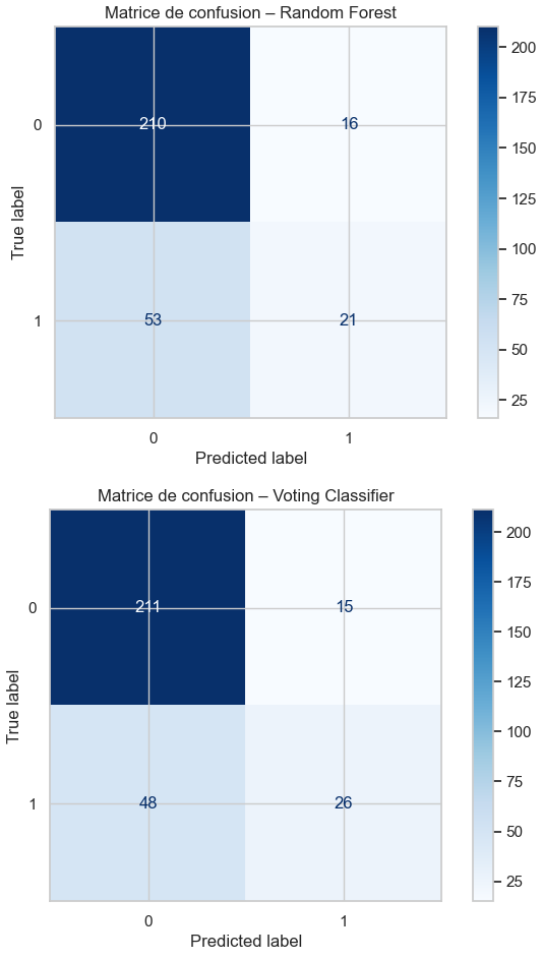

Figure 6: Matrices de confusion : Random Forest et Voting Classifier

C.2 Courbes ROC comparées

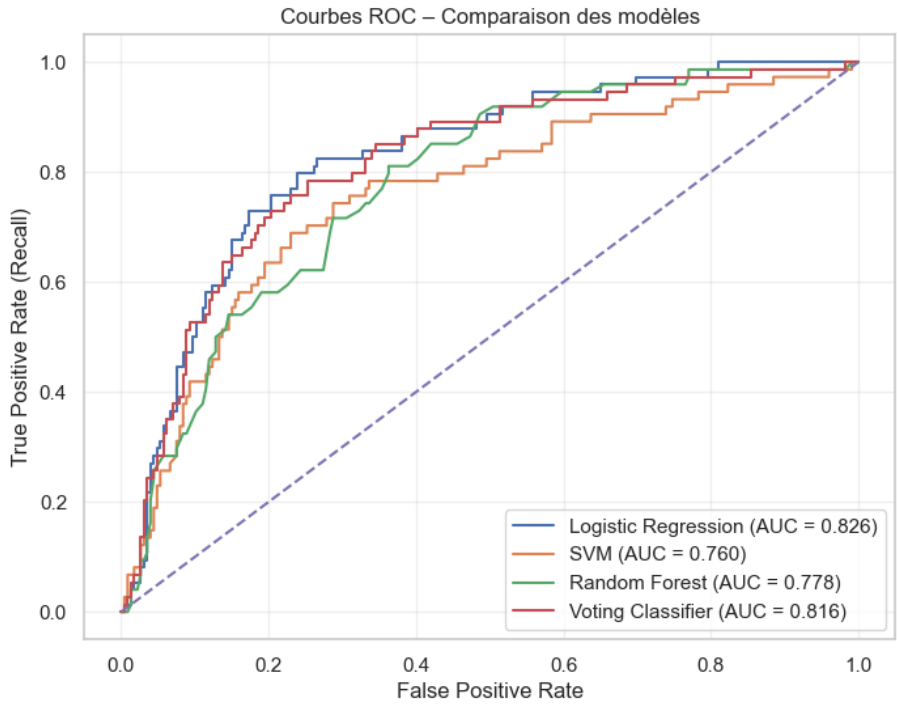

Figure 7: Courbes ROC des différents modèles

D Analyse de seuil

D.1 Analyse de seuil du Random Forest

Table 3: Impact du seuil de décision sur les performances

Seuil	Précision	Rappel	F1-score	FN / FP
0.5	0.553	0.284	0.375	53 / 17
0.4	0.494	0.581	0.534	31 / 44
0.3	0.401	0.824	0.540	13 / 91

E Interprétation du modèle par SHAP

E.1 Importance globale des variables

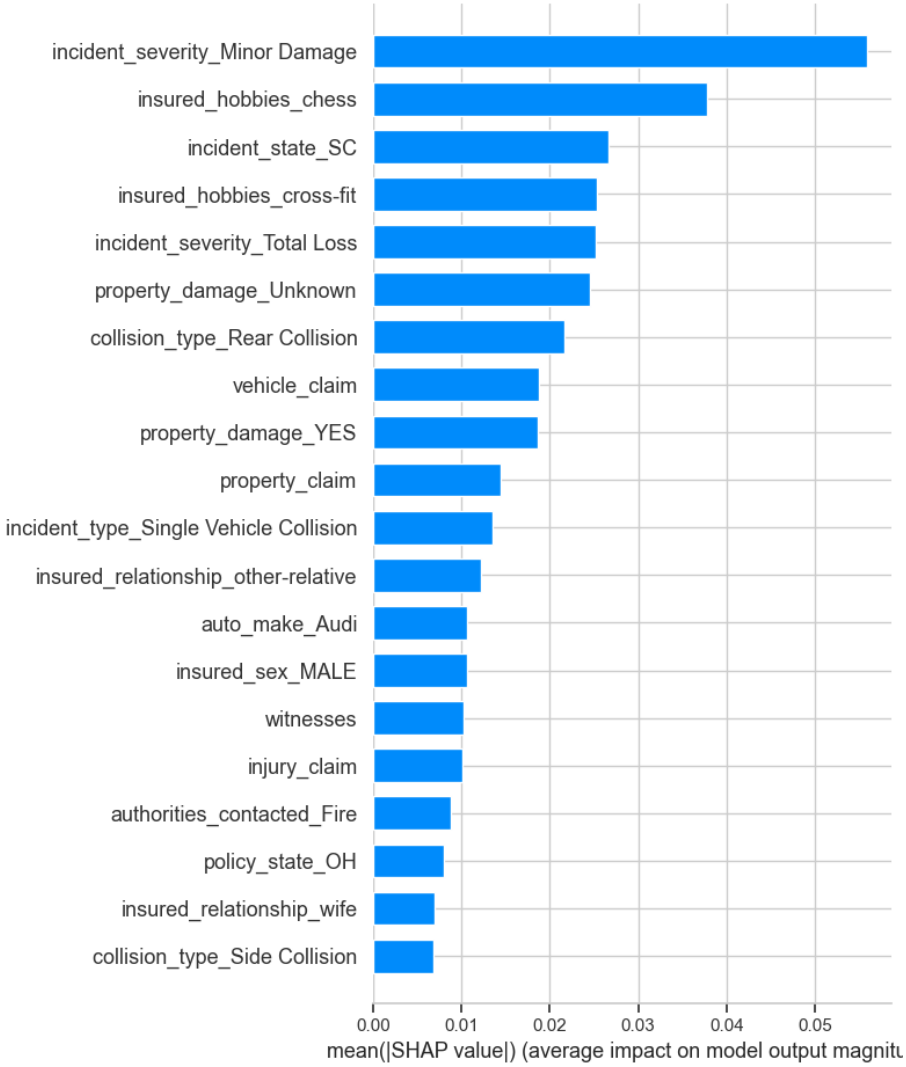

Figure 8: Importance globale des variables selon SHAP

E.2 Analyse locale des contributions

Figure 9: Distribution des contributions SHAP (beeswarm)