**Mémoire présenté le :**

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires**

Par : Khouloud BARHOUMI

Titre **Modélisation de la fraude en assurance IARD : Application à l'assurance automobile**

Confidentialité : ☐ NON ■ OUI (Durée : ☐ 1 an ■ 5 ans)

*Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus*

*Membre présents du jury de l'Institut des Actuaires* signature

Mme Sophie DECUPERE

M. Frédéric SCHWACH

M. Romain NOBIS

*Membres présents du jury de l'ISFA*

M. Stéphane LOISEL

M. Aurélien COULOUMY

*Entreprise :*

Nom : **Covéa**

*Signature :*

*Directeur de mémoire en entreprise :*

*Nom : Mathieu Bonelli*

*Signature :*

*Invité :*

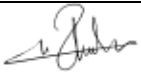
*Nom :*

*Signature :*

***Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels*** *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat



# Modélisation de la fraude en assurance IARD : Application à l'assurance automobile

Khouloud Barhoumi

27 mai 2020

**Résumé**

La lutte contre la fraude est devenue aujourd'hui un enjeu majeur dans la gestion des risques des compagnies d'assurance.
Le coût total des fraudes à l'assurance IARD est estimé à plus de 2.5 milliards d'euros. La détection de la fraude constitue alors une problématique de premier plan pour faire face aux exigences en matière de modélisation et prédiction dûes au faible ratio de fraudes connues dans les échantillons.
La stratégie de la lutte contre la fraude de Covéa s'appuie historiquement sur l'expertise des centres de gestion et des cellules spécialisées. Cette approche traditionnelle, basée sur le développement de scénarii et règles métiers, a montré ses limites du fait de sa faible vélocité et de son manque d'adaptabilité. Afin de pallier à ces défauts, les assureurs portent désormais un intérêt particulier aux algorithmes d'apprentissage automatique dans l'objectif de disposer des systèmes de détection de fraude automatisés et industrialisables.
Ce mémoire présente un éventail de ces techniques d'apprentissage automatique utilisables dans le cadre de la détection des réclamations frauduleuses en assurance automobile. Ces modèles sont calibrés en utilisant différentes bases de données incluant des cas de fraudes identifiées et avérées. Une étude est également menée afin d'améliorer les performances de ces modèles par l'intermédiaire de la sélection, le traitement et l'identification des variables significativement discriminantes.

Mots clés : **Assurance automobile, Fraude, Apprentissage automatique, Apprentissage supervisé**

## Abstract

Nowadays, the fight against fraud has become a major issue in risk management in insurance companies.
The total cost of P&C insurance fraud is estimated at more than 2.5 billion euros. The fraud detection is therefore a challenging problem in terms of modeling and prediction requirements due to the low known fraud ratio in the samples.
Covéa's anti-fraud strategy has historically relied on the expertise of management centers and specialized cells. This traditional approach, based on the development of scenarios and business rules, has shown its limits due to its low velocity and its lack of adaptability. In order to overcome these shortcomings, insurers are now taking particular interest in machine learning algorithms with the aim of having automated and industrializable fraud detection systems.
This thesis presents a range of these machine learning techniques that can be used in the context of detection of fraudulent automobile insurance claims. These models are calibrated using different databases including cases of identified and proven fraud. A study is also conducted to improve the performance of these models through the selection, preprocessing and identification of significantly discriminating variables.



Key words : **Automobile Insurance**, **Fraud**, **Machine Learning**, **Supervised Learning**

# Remerciements

La réalisation de ce projet a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma reconnaissance.

D'abord, je tiens à exprimer ma gratitude envers mon encadrant M. Mathieu Bonelli Manager Big Data et Socle de données *Covéa* pour ses conseils, son dévouement et son suivi du projet.

Je tiens à remercier mon sponsor métier *Covéa* M. Alexis Durand et notre experte du pôle lutte anti-fraude Mme Karine Torres pour leurs assistance.

Je remercie également tous les collaborateurs *Covéa* des directions Sinistres, Actuariat, Data Science et Informatique pour leur disponibilité et leur aide.

Mes sincères remerciements sont adressés aussi à l'équipe du Centre de solutions Big Data et socles de données Covéa pour son accueil chaleureux et les précieux conseils de ses collaborateurs tout au long de mon stage.

# Table des matières

# Liste des algorithmes

# Liste des tableaux

# Table des figures

# Introduction

La fraude est un fléau pour les compagnies d'assurance. Fausses déclarations à la souscription, sinistres fictifs, exagération des montants des dommages... l'ensemble de ces actes intentionnels de l'assuré dans le but de dégager un profit illicite du contrat d'assurance représente un coût estimé à 2.5 milliards d'euros selon l'Agence de la Lutte contre la Fraude à l'Assurance *ALFA*[1], soit 5% des primes *IARD*[2] collectées chaque année. Les 5% de fraude vont à l'encontre du principe de mutualisation car cela impacte la tarification de l'ensemble des assurés et dégradent la rentabilité de l'assureur dans un secteur d'activité à faible marge.

En matière de lutte contre la fraude, les assureurs s'appuient, d'une part, sur la sensibilisation de toutes les parties prenantes internes (services de souscription, gestionnaires de sinistres, experts en assurance et en juridique) et des collaborateurs externes via des collaborations avec des autorités et des entités privées. Vu le coût important de la fraude, les assureurs cherchent à renforcer la détection des comportements frauduleux et posséder un dispositif robuste de lutte contre la fraude.

Ce dispositif repose sur quatre piliers :

— La connaissance de l'exposition au risque de la fraude

— La disposition d'un processus de prévention et de contrôle

— L'industrialisation de la détection en s'appuyant sur l'analyse de données

— L'optimisation des modèles de détection de la fraude.

1. Agence pour la Lutte contre la Fraude à l'Assurance
2. Incendie, Automobile et Risques Divers

L'analyse classique des données par les référents opérationnels fraude permet de créer des indicateurs appelés des *règles métiers* afin de remonter des alertes de suspicion. Les *règles métiers* sont considérées comme une réplication des cas de fraude déjà connus. Cette approche traditionnelle adoptée par les assureurs se confronte à la grande quantité de données à analyser, ce qui se traduit généralement par un temps de production des alertes de suspicion non compatible avec l'objectif de traitement d'un sinistre simple en moins d'une semaine. Avec l'apparition des nouvelles technologies notamment le *Big Data* et la *Data Science*, l'utilisation des techniques de *Machine Learning* peut pallier à ce problème. *Covéa* cherche à mettre en place un système automatisé de détection des comportements frauduleux en utilisant les entrepôts de données que le groupe possède et en faisant appel à des techniques prédictives. Cette approche est considérée comme un complément du modèle déterministe traditionnel et vise à identifier les cas de fraudes cachés afin d'agir dans les meilleurs délais.

Les méthodes de *Machine Learning*, qu'elles soient supervisées ou non supervisées, permettent d'identifier les corrélations complexes entre un grand nombre de variables et à détecter les signaux faibles de la fraude. En effet, les méthodes supervisées qui se basent sur un référentiel des fraudes identifiées sont efficaces pour repérer les typologies des fraudeurs et confirmer les assertions des experts et des gestionnaires de sinistres. Quant aux méthodes non supervisées, elles consistent à identifier les cas d'anomalie et attribuer une probabilité de frauder. Cette approche est intéressante dans le cas où le nombre de dossiers frauduleux ne représente qu'une faible proportion des observations.

En littérature, certains travaux montrent l'efficacité de quelques modèles testés dans le cadre de la recherche de fraude à l'assurance. En apprentissage supervisé, les méthodes basées sur *les arbres de décision* font l'objet de plusieurs études, notamment *les forêts aléatoires* et les méthodes de *boosting*. En apprentissage non supervisé, les méthodes les plus appropriées sont les modèles de détection d'anomalie. Elles visent à classifier les observations entre normales et aberrantes ayant un comportement marginal.

Ce mémoire s'inscrit dans le cadre d'un projet de recherche et développement initié par *Covéa* dont l'objectif est d'évaluer les capacités internes à identifier les fortes suspicions. Parallèlement, *Covéa* est en train de créer des entrepôts de données spécifiques au traitement de la fraude.

Notre travail sera structuré en quatre parties :

— Dans le premier chapitre, nous présenterons les formes de fraude à l'assurance *IARD* ainsi que les sanctions énoncées par le code des assureurs.

— Dans le deuxième chapitre, nous ferons le point autour de différentes méthodes de *Machine Learning* vues en littérature permettant de détecter efficacement les cas de fraude.

— Le troisième chapitre sera consacré à la modélisation de la fraude à travers l'application des algorithmes sur un portefeuille des sinistres automobile d'une des marques du groupe *Covéa*.

— L'objectif du dernier chapitre est d'analyser et de comparer les résultats des différents modèles et également de présenter les pistes d'améliorations envisageables afin d'avoir le modèle le plus performant.

Chapitre 1

# La fraude à l'assurance

## 1.1 Contexte

Dans ce chapitre, nous allons présenter les différents concepts liés à la fraude en assurance automobile, qui est la priorité chez covéa. La principale raison pour laquelle les assureurs s'intéressent au traitement de la fraude en assurance automobile est la disponibilité des données. En effet, au niveau du marché, c'est dans cette branche que le volume de fraude est le plus important et que les données sont les plus standardisées. L'objectif de mettre en place un dispositif d'automatisation d'aide à la détection de fraude y devient stratégique, en traitant un grand volume de données plus rapidement. Chez *Covéa*, les premiers projets d'automatisation de la détection de fraude sont menés dans cette branche.

### 1.1.1 Assurance automobile

L'assurance automobile en France est obligatoire depuis 1958 et est destinée aux véhicules terrestres à moteur *VTM*. Il s'agit de « tout véhicule pourvu d'un moteur destiné à circuler sur le sol sans être lié à une voie ferrée. En plus des automobiles et deux-roues, les engins de chantier, les machines agricoles, les remorques et semi-remorques font donc partie des véhicules terrestres à moteur. »

Le marché de l'assurance automobile représente environ 10% du marché de l'assurance en France. Cette branche d'assurance représente également 40% du total des cotisations en assurance aux biens.

Elle propose plusieurs garanties obligatoires et facultatives :

**La garantie responsabilité civile** Cette garantie obligatoire permet d'indemniser les dommages matériels et corporels causés par le conducteur du véhicule à un tiers, que le conducteur soit déclaré ou non. Le non respect de cette garantie est constitutif d'un délit.

**La garantie vol** permet à l'assuré d'être indemnisé si le véhicule est volé. Cette indemnité est égale à la valeur du véhicule le jour du vol ou à une valeur précisée dans le contrat d'assurance. En annexe à la garantie vol, d'autres garanties peuvent être couvertes dans le contrat comme les tentatives de vol, le vol des accessoires du véhicule et le vandalisme.

**La garantie incendie** permet de couvrir les cas d'incendie et d'explosion.

**La garantie bris de glace** permet de couvrir les dommages subis par les parties vitrées du véhicule comme le pare-brise, les glaces latérales, la lunette arrière, les blocs optiques de phares et les rétroviseurs.

**La garantie dommages** concerne le cas de collision avec un piéton, un animal, un objet fixe ou un autre véhicule permettant ainsi de couvrir les dommages causés au véhicule.

**La garantie catastrophes naturelles** permet d'indemniser l'assuré pour les dégâts subis par le véhicule à cause d'une catastrophe naturelle (inondation, avalanche, tremblement de terre...).

**La garantie assistance** consiste au dépannage et du rapatriement de l'assuré en cas de pannes, d'accidents de la route, de pannes de carburants et de vol.

## 1.2 La fraude à l'assurance

Dans ce projet, nous nous intéressons à la fraude à l'assurance qui concerne la fraude d'un assuré envers son assureur. Selon l'Agence pour la Lutte contre la Fraude à l'Assurance ALFA [1], la fraude est définie comme « *un acte volontaire permettant de tirer un profit illégitime d'un contrat d'assurance* ». La fraude affecte toutes les branches de l'assurance. Un individu peut frauder de diverses manières. La fraude peut concerner le contrat

1. Cet organisme sera présenté dans la section suivante.

d'assurance au niveau de la souscription ou bien le sinistre lors de son survenance. Le code des assurances sanctionne de manière spécifique et sévère les cas de fraude à l'assurance.

### 1.2.1 Les types de fraude

Il existe de nombreuses formes de fraude qui peuvent être classifiées selon la nature des différents scénarios.

#### 1.2.1.1 La fraude à la souscription

L'objectif principal de la fraude à la souscription est de faire baisser la prime de l'assurance. Il s'agit de donner des informations et des renseignements erronés lors de la souscription du contrat d'assurance. Le souscripteur dissimule des éléments aggravants du risque dans le but d'avoir un tarif plus réduit.

#### 1.2.1.2 La fraude lors de la survenance d'un sinistre

Il existe plusieurs manifestations de la fraude lors de la survenance d'un sinistre.

**La fraude planifiée**

Il s'agit d'un faux sinistre organisé ou d'une fausse déclaration d'un sinistre portant sur les causes et les circonstances de survenance. Par exemple, le sinistré peut faire disparaitre son véhicule (généralement invendable à cause de son mauvais état) et déclare un sinistre pour profiter de l'indemnisation de la garantie vol. Les incendies volontaires appartiennent aussi à cette forme de fraude.

**La fraude opportuniste**

Ce type de fraude concerne la déclaration du sinistre. Le sinistré profite de la survenance d'un sinistre pour se faire indemniser des dommages antérieurs en modifiant les circonstances de la survenance du sinistre ou bien la garantie.

**La fraude par exagération du montant de sinistre**

Ce procédé est très courant. Il consiste à falsifier les pièces justificatives telles que les factures afin d'avoir une estimation du montant de réparation

supérieure au coût réel et d'obtenir ainsi un bénéfice.

### 1.2.2 Les sanctions

Afin de prévenir et dissuader la fraude, le code des assurances prévoit des sanctions pour les différentes formes de cette pratique illégale à travers les articles L.113-8 et L.113-9.

#### La fraude à la souscription

Le code de l'assurance impose à l'assuré, à travers l'article L.113-2, de répondre exactement aux questions posées par l'assureur, notamment dans le formulaire du risque lors de la conclusion du contrat. L'identité du conducteur, l'usage du véhicule, etc. sont des renseignements indispensables à l'assureur pour tarifer le risque et calculer les cotisations.

Il appartient à l'assureur de faire la preuve de l'inexactitude dans la déclaration du risque. L'assuré est présumé de bonne foi sauf preuve du contraire.

Lorsque la mauvaise foi est établie, l'article L.113-8 est appliqué et l'assureur peut donc demander en justice la nullité du contrat.

En effet, l'article L.113-8 du code des assurances prévoit :

≪ *Indépendamment des causes ordinaires de nullité, et sous réserve des dispositions de l'article L. 132-26, le contrat d'assurance est nul en cas de réticence ou de fausse déclaration intentionnelle de la part de l'assuré, quand cette réticence ou cette fausse déclaration change l'objet du risque ou en diminue l'opinion pour l'assureur, alors même que le risque omis ou dénaturé par l'assuré a été sans influence sur le sinistre. Les primes payées demeurent acquises à l'assureur, qui a le droit au paiement de toutes primes échues à titre de dommages et intérêts.* ≫

Dans le cas de la déclaration inexacte du risque, dénaturation du risque ou modification de l'appréciation du risque, sans mauvaise foi, découverte par l'assureur avant un sinistre, l'assureur a le droit soit de maintenir le contrat en augmentant la prime (acceptée par l'assuré), soit de résilier le contrat. Si l'erreur commise par l'assuré dans la déclaration du risque est découverte par l'assureur suite à un sinistre, la Règle Proportionnelle de Prime est appliquée.

L'article L.113-9 du code des assurances indique :

*« Dans le cas où la constatation n'a lieu qu'après un sinistre, l'indemnité est réduite en proportion du taux des primes payées par rapport au taux des primes qui auraient été dues, si les risques avaient été complètement et exactement déclarés. »*

Le montant de l'indemnité versée se calcule simplement par :

$$\text{Montant de l'indemnité versée} = \text{Montant des dommages évalués} \times \frac{\text{Prime payée}}{\text{Prime due}}$$

**La fraude lors de la survenance d'un sinistre**

Le code des assurances énonce, pour les fraudes aux sinistres, que :

> *« L'assuré qui a fait de mauvaise foi une déclaration inexacte relative au sinistre est déchu du bénéfice de l'assurance. »* [4]
>
> — article L. 172-28 du code des assurances.
>
> *« Les pertes et les dommages occasionnés par des cas fortuits ou causés par la faute de l'assuré sont à la charge de l'assureur, sauf exclusion formelle et limitée contenue dans la police.*
>
> *Toutefois, l'assureur ne répond pas des pertes et dommages provenant d'une faute intentionnelle ou dolosive de l'assuré. »*[4]
>
> — article L. 113-1 du code des assurances.
>
> L'assureur peut donc refuser l'indemnisation du sinistre. Et si l'indemnité a été payée avant que la fraude soit avérée, l'assureur peut demander le remboursement des sommes versées.

### 1.2.3 La fraude en chiffres

**L'Agence pour la Lutte contre la Fraude ALFA**

L'Agence pour la Lutte contre la Fraude à l'Assurance *ALFA* est créée en 1989. Il s'agit d'une association dont le but est de promouvoir la lutte contre la fraude en assurance et de doter les compagnies d'assurance d'une structure opérationnelle anti-fraude.

En 2018, les données collectées par ALFA pour la branche $IARD$[2] concernent près de 80% de part de marché. 36 887 sinistres frauduleux avérés ont été identifiés par les assureurs suite à une analyse portée sur 76,59% du marché

2. Incendie, auto et risques divers

*IARD*. L'enjeu financier est important et est estimé à 431 millions d'euros dont les économies réalisées représentent 324 millions d'euros.

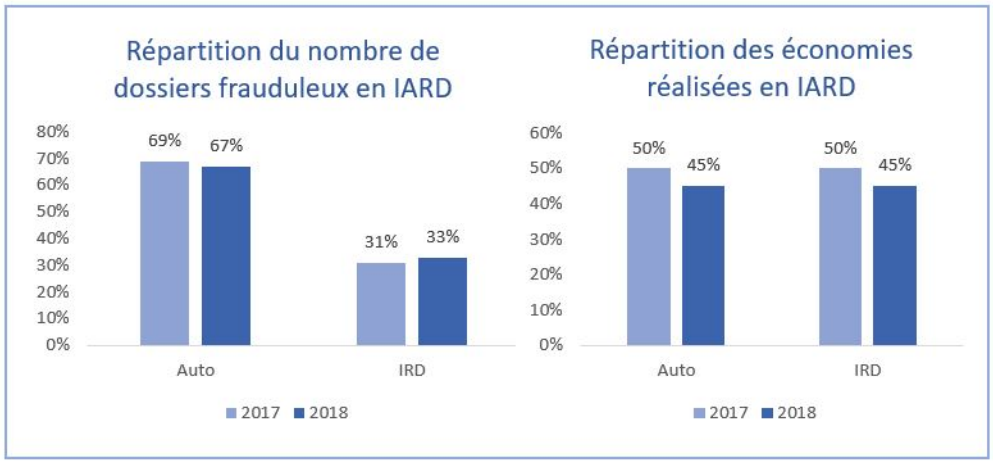
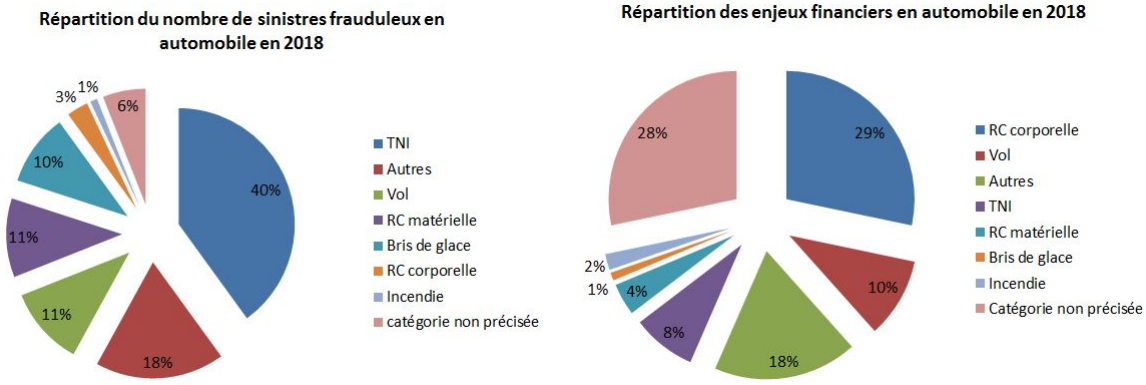



FIGURE 1.1 – Statistiques des fraudes avérées par le marché en 2018 (Source : ALFA)

Pour la branche automobile, les dossiers frauduleux représentent 67% du nombre total des dossiers frauduleux en IARD, soit 24 842 sinistres. Les économies réalisées sur la branche automobile sont estimées à 145 millions d'euros.




TNI : Tiers Non Identifié

FIGURE 1.2 – Répartition du nombre de dossiers frauduleux en automobile en 2018 ( Source : ALFA)

### 1.2.4 Impact de la fraude : Enjeux techniques et stratégiques

Pour sa part, Solvabilité 2 oblige les assureurs à prendre un certain nombre de mesures dont : l'identification, l'évaluation et la gestion des risques opérationnels qui ont un effet potentiel sur le capital. Le risque de fraude est considéré comme intégré dans la gestion des risques opérationnels.

La détection de la fraude à l'assurance est appréhendé comme étant un problème d'asymétrie d'information entre l'assuré et l'assureur pouvant engendrer des phénomènes de risque moral et de sélection adverse.

## 1.3 Les actions de prévention dans les compagnies d'assurance

L'assureur est le premier à être impacté par la fraude et à subir les pertes engendrées par cet acte. De ce fait, il est primordial pour l'assureur de mettre en place des actions de sensibilisation, de prévention et un dispositif de détection de fraude. L'impact existe également pour les clients. L'assureur doit préserver l'intérêt de ses clients honnêtes afin de ne pas leur faire supporter le surcout de l'indemnisation indue de clients malhonnêtes.

Dans cette optique, l'assureur a recours à des organismes et des fichiers professionnels afin de vérifier les informations fournies par les clients lors de la déclaration du risque et limiter ainsi la prise en charge des sinistres frauduleux.

D'autres sources s'ajoutent à l'agence pour la lutte contre la fraude *ALFA*, nous citons :

#### AGIRA

C'est l'Association pour la Gestion des Informations sur le Risque en Assurance. C'est une association créée par la Fédération Française des Sociétés d'Assurance *FFSA* et par le Groupement des entreprises mutuelles d'assurance *Gema*. Elle fournit un fichier des résiliations automobiles dans le but de conserver une trace des antécédents d'assurance d'une personne. Ce fichier comporte l'état des sinistres sur les 5 dernières années et également le motif de la suspension ou la résiliation du contrat d'assurance. Ce fichier est alimenté par l'ensemble des compagnies d'assurance automobile. Il permet, lors de l'enregistrement d'un nouveau risque, de vérifier les informations fournies par le client.

**ARGOS**

Il s'agit d'un groupement des assureurs français pour l'identification, la recherche et la récupération des véhicules et autres biens mobiliers déclarés volés en vue d'indemnisation.

**SIDEXA**

Créée en 1981, *Sidexa* est membre du groupe *Solera* et leur objectif est d'être la plateforme digitale de la gestion des risques et des biens. La base de données *Sidexa* constitue une référence de nombreux professionnels de l'automobile dont les experts pour calculer méthodiquement le montant de réparations et établir instantanément les rapports d'expertise, devis et valeurs de véhicule.

Ces fichiers et sources permettent aux assureurs de vérifier les informations fournies lors de la souscription du contrat et de déterminer ainsi les fausses déclarations intentionnelles.

Lors de la survenance du sinistre, la mission de la détection de la fraude est confiée aux experts et gestionnaires de sinistres. Ils sont habilités à déceler les premiers indices signalant une éventuelle tentative de fraude.

Certains assureurs font également appel à des prestataires externes spécialisés dans la détection de fraude afin de pouvoir agir dans les meilleurs délais suite à une alerte.

Dans cette optique, et pour faciliter la tâche de détection de la fraude et des manifestations variées aux gestionnaires du sinistre et aux experts, les assureurs ont recours aux nouvelles méthodes et technologies permettant de générer automatiquement des alertes de suspicion.

Chapitre 2

# Les algorithmes de détection de fraude

La détection de fraude à l'assurance est un enjeu majeur pour les compagnies d'assurance mais se montre complexe à traiter compte tenu de la diversité des types de fraude et du nombre relativement faible de fraudes démontrées dans des échantillons types. Les techniques d'apprentissage automatique permettent d'optimiser la précision prédictive et ainsi d'améliorer la qualité des alertes analysées par les experts anti-fraude des compagnies d'assurance.

L'apprentissage automatique[1] *( également nomé Machine Learning*[2]) est un domaine d'étude de l'intelligence artificielle qui s'appuie sur des algorithmes et des techniques statistiques pour permettre aux systèmes informatiques d'apprendre à partir des données.

L'apprentissage automatique est généralement constitué de deux phases. La première correspond à la construction et l'estimation d'un modèle à partir des données. Cette phase est appelée la phase d'apprentissage ou bien l'entrainement. La seconde phase consiste à la validation du modèle. En effet, le modèle déterminé à partir de la première phase est ensuite appliqué sur des nouvelles données pour prédire les résultats correspondants.

Deux catégories peuvent être distinguées dans l'apprentissage automatique selon le type des données utilisées dans la phase d'apprentissage, et

1. L'idée a été concrétisée grâce aux travaux de Alan Turing concernant la ≪ machine universelle ≫ en 1936 et ≪ l'ordinateur et l'intelligence ≫ en 1950.

2. Ce terme a été utilisé pour la première fois en 1959 par Arthur Samuel, un informaticien américain

sont l'apprentissage supervisé et l'apprentissage non supervisé.

## 2.1 Apprentissage statistique

### 2.1.1 Apprentissage supervisé

Dans l'apprentissage supervisé, les observations sont étiquetées ou annotées. Il s'agit d'un problème de régression si la variable cible est une variable quantitative (continue) ou bien un problème de classement si la variable cible est une variable qualitative (le nombre des valeurs de sortie est fini, par exemple des classes ou des catégories).

L'échantillon des données labellisées (annotées) est considéré comme une base d'apprentissage qui sert à entrainer le modèle et le généraliser, c'est-à-dire, apprendre au modèle à faire des prédictions correctes sur de nouvelles données non annotées à partir des données explicatives correspondantes.

Il s'agit ainsi de définir une liaison fonctionnelle sous-jacente ( en anglais, *underlying concept*) entre la variable cible et les variables explicatives de la forme

$$Y = f(X, \alpha)$$

avec
X représente les variables explicatives
Y est la variable cible
et $f(.)$ est le modèle de prédiction de paramètre $\alpha$.

Le problème de détection de fraude peut être résolu à l'aide des techniques d'apprentissage supervisé. Il s'agit ainsi d'un problème de classement et consiste à affecter un individu à la classe des fraudeurs ou bien des non fraudeurs.

### 2.1.2 Apprentissage non supervisé

A la différence de l'apprentissage supervisé, le contexte non supervisé correspond aux algorithmes qui doivent opérés à partir des observations non annotées. Un algorithme d'apprentissage non supervisé cherche automatiquement à associer des catégories aux données et ainsi à les classer. En effet, l'algorithme cherche à maximiser l'homogénéité des données et à parvenir ainsi à former des groupes de données distincts.

Il existe plusieurs algorithmes d'apprentissage non supervisé :

— Les méthodes de partitionnement des données *Clustering* (par exemple *K-means*, le regroupement hiérarchique),

— Les méthodes de réduction de dimension (par exemple l'analyse en composantes principales $ACP$)

— Les méthodes de détection d'anomalies. Cette approche consiste à identifier les évènements rares qui diffèrent significativement de la majorité des observations. Ces observations sont appelées des anomalies ou bien des valeurs aberrantes (*outliers*). Dans le cas de détection de fraude, les observations aberrantes peuvent soulever des suspicions.

## 2.2 La régression logistique

La régression logistique est un modèle mathématique simple qui permet de prédire une variable cible binaire à partir des variables explicatives quantitatives ou bien qualitatives.

La régression logistique est un cas particulier du modèle linéaire généralisé dans le but de modéliser la probabilité d'appartenance de la variable cible à une classe en fonction des variables explicatives.

### Explication du modèle

Soit $X = (X_1, X_2, ..., X_p)$ les variables explicatives et Y la variable cible binaire (variable à expliquer). Les modalités possibles de Y sont {0,1}.

Notons $p(X|Y)$ la distribution conditionnelle des X sachant la valeur prises par Y, et $p(Y = 1|X)$ (respectivement $p(Y = 0|X)$) est la probabilité a postériori d'obtenir la modalité 1 de Y ( respectivement 0) sachant les valeurs prises par X.

En appliquant le théorème de Bayes, la probabilité $P(Y = 1|X)$ s'écrit :

$P(Y = 1|X) = \frac{P(Y=1)*P(X|Y=1)}{P(X)}$
et respectivement $P(Y = 0|X) = \frac{P(Y=0)*P(X|Y=0)}{P(X)}$

et ainsi,

$$\frac{P(Y = 1|X)}{P(Y = 0|X)} = \frac{P(Y = 1)}{P(Y = 0)} * \frac{P(X|Y = 1)}{P(X|Y = 0)}$$

Dans le but d'estimer ce rapport de probabilité, la régression logistique est fondée sur l'hypothèse fondamentale suivante

$$\ln[\frac{P(X|Y=1)}{P(X|Y=0)}] = b_0 + b_1X_1 + ... + b_pX_p$$

Y suit une loi de Bernoulli et donc $P(Y=1|X) = \mathbb{E}(Y|X) = \pi$, $\pi$ varie entre 0 et 1.

Une manière différente de décrire la régression logistique est d'introduire la fonction LOGIT [3] :

$$ln[\frac{\pi}{1-\pi}] = a_0 + a_1X_1 + ... + a_pX_p$$

et donc

$$\pi = \frac{\exp(a_0 + a_1X_1 + ... + a_pX_p)}{1 + \exp(a_0 + a_1X_1 + ... + a_pX_p)}$$

$$\pi = \frac{1}{1 + \exp(-(a_0 + a_1X_1 + ... + a_pX_p))}$$

Les paramètres de la régression logistique sont estimés par la méthode du maximum de vraisemblance. L'estimation des paramètres n'est pas un exercice évident, plusieurs algorithmes sont utilisés mais le plus connu est la méthode de Newton-Raphson.

## Évaluation du modèle

Parmi les mesures permettant d'évaluer la qualité d'ajustement d'un modèle, nous trouvons *l'AIC*.

### AIC

C'est le critère d'information d'Akaike. Il permet de mesurer la qualité d'un modèle statistique. C'est un critère de pénalisation des modèles en fonction du nombre de paramètres. Il est défini en fonction du nombre de paramètres k dans le modèle et le maximum de la fonction de vraisemblance L permettant ainsi d'avoir un compromis entre la complexité du modèle et la qualité d'ajustement. Il s'écrit comme

$$AIC = 2k - 2ln(L)$$

3. Voir détails en annexe 1.

Le meilleur modèle est obtenu en choisissant la valeur la plus faible *d'AIC*.

## 2.3 Les forêts aléatoires

Les forêts aléatoires ou en anglais *Random Forest* sont une méthode d'apprentissage automatique introduite par Breiman [4] en 2001. C'est une technique d'agrégation d'un ensemble d'arbres de décision *CART*[5], et ainsi elle est plus performante que les simples arbres. Cet algorithme peut être utilisé pour des problèmes de régression ou bien de classement.

### 2.3.1 Algorithme

Les forêts aléatoires sont construites en agrégeant les arbres *CART*. Ce modèle est d'autant plus performant que la corrélation entre les arbres est faible. [ Breiman 2001]

Nous avons en entrée x comme l'observation à prévoir et p variables explicatives.

**Algorithme 2.1** Les forêts aléatoires

Tirer avec remise dans la base d'apprentissage B échantillons $z_i$ , $i = 1, ..., B$ (chaque échantillon ayant n points $z = \{(x_1, y_1), ..., (x_n, y_n)\}$)
Pour $i = 1, ..., B$ :

- Tirer un échantillon $z_i$
- Construire un arbre de décision $G_i(x)$ selon une condition de tirer aléatoirement un sous-ensemble des attributs (choisir q variables parmi p variables explicatives) lors du découpage d'un nœud (split)
- Choisir le meilleur découpage dans ce sous-ensemble

Agrégation par vote G(x)= Vote majoritaire $(G_1(x), .., G_B(x))$ (agrégation par la moyenne en cas de régression)

Par défaut, le choix de q est fait en prenant $q = \sqrt{p}$ ($q = p/3$ en régression). Ce nombre est fixé et est identique pour tous les arbres.

4. Leo Breiman était un statisticien américain à l'université de Californie et connu par ses travaux dans le domaine d'apprentissage automatique notamment sur les arbres de régression et de classification.

5. Cette méthode de Machine Learning est expliquée en annexe.

Les forêts aléatoires permettent d'avoir des arbres de décision moins corrélés. En effet, le tirage aléatoire des variables explicatives à chaque nœud permet d'aboutir à des arbres indépendants car les arbres sont entrainés sur un ensemble différent de variables explicatives (aléa du choix des attributs) et sur des échantillons différents (*Boostrap*).[Breiman et Cutler 2004]

L'application de *Random Forest* sur les données propose plusieurs sorties. Parmi ces sorties, nous trouvons, l'erreur *Out-of-bag* et l'importance des variables.

#### Erreur Out-of-bag

Pour prévenir le sur-apprentissage, l'estimation de l'erreur *Out-of-bag* est nécessaire. Il s'agit d'une estimation de l'erreur de prédiction en profitant de l'information fournie par les estimateurs agrégés. Cette erreur permet de contrôler le nombre d'arbres B (défini comme *ntree* dans les fonctions de *random forest*).

Soit un individu $(x_i, y_i)$ de l'échantillon d'apprentissage, nous construisons des arbres sur les échantillons *boostrap* ne contenant pas cet individu. On dit alors que cet individu est *out-of-bag*. Nous faisons une prédiction de la classe de cet individu en agrégeant les prédictions des arbres. Notons $\hat{y}_i$ la prédiction de $y_i$. Cette opération est appliquée sur toutes les observations de la base. L'erreur est calculée en évaluant la proportion d'individus mal classés $\frac{1}{B}\sum_{i=1}^{B} \mathbf{1}_{\hat{y}_i \neq y_i}$ (C'est l'erreur quadratique moyenne en régression).

### 2.3.2 L'importance des variables

Dans le cas de données de très grandes dimensions, il est utile de déterminer les variables explicatives ayant un rôle important dans le modèle et permettant d'obtenir une bonne prédiction. La Forêt aléatoire fournit un classement des variables fondé sur leurs importance vis-à-vis la variable cible.

#### Méthode des permutations de Breiman

L'évaluation de l'importance des variables explicatives peut se faire à partir des permutations. Cette méthode a été proposée par Breiman en 2001. Il s'agit de permuter les valeurs d'une variable donnée dans l'échantillon *OOB* (*out-of-bag*) et évaluer l'erreur *out-of-bag* de l'arbre dans la forêt. Concrètement, pour chaque arbre i, $i \in \{1, ..., B\}$, l'erreur *OOB* est calculée et correspond à la proportion des observations mal classées dans l'échantillon *out-of-bag*. Ensuite, pour une variable fixée $X_j$, $j \in \{1, ..., p\}$, l'erreur est

réévaluée après la permutation des valeurs de la variable $X_j$. La mesure de l'influence de chaque variable $X_j$ et pour chaque arbre i correspond à la différence entre les erreurs calculées avant et après la permutation des valeurs de la variable. Cette procédure est effectuée pour toutes les variables et pour chaque arbre de la forêt. Ainsi, l'importance d'une variable donnée correspond à la moyenne des différences des erreurs obtenues pour tous les arbres. Plus l'erreur augmente, plus la variable est importante. Cette méthode est généralement appelée *Mean Decrease accuracy.*

#### Diminution moyenne de l'impureté

Appelée également diminution moyenne *Gini* (en anglais *mean decrease Gini*), c'est une approche proposée par Breiman pour évaluer l'influence de chaque variable à partir des arbres *CART*. Cette méthode se base sur l'évaluation de l'indice de *Gini*[6] dans le calcul de l'erreur de prédiction dans l'arbre. Pour chaque variable $X_j$, $j \in \{1, ..., p\}$, cette mesure consiste à calculer la moyenne sur tous les arbres, de la somme des diminutions de l'impureté de *Gini* pour tous les nœuds lorsque cette variable $X_j$ intervient dans la découpe pondérées par le nombre d'observations dans le nœud. Plus la *mean decrease Gini* est élevée, plus la variable est importante.

## 2.4 Extreme Gradient Boosting : XGBoost

### 2.4.1 Principe du Boosting

Le *Boosting* est un domaine d'apprentissage statistique qui repose sur l'agrégation récursive d'une famille de modèles par un vote majoritaire (par une moyenne pondérée en régression). En effet, chaque modèle représente une version adaptative du précédent en attribuant un poids supplémentaire aux observations mal classées ou mal prédites.

L'idée générale a été développée par Freund et Schapure en 1996 en présentant l'algorithme original *adaboost* (*Adaptative Boosting*) dans le cadre d'un problème de classification binaire. Des adaptations de cet algorithme ont été publiées pour des problèmes de régression et de classification multi-classes.

6. Voir annexe 2.

### Algorithme de base : Adaboost

C'est l'algorithme le plus populaire et la version originale du *boosting*. Il est conçu pour un problème de classification binaire, à deux classes dont les valeurs sont dans $\{-1, 1\}$. Soit $x$ la variable à prévoir et $z = \{(x_1, y_1), ..., (x_n, y_n)\}$ l'échantillon d'apprentissage. L'algorithme commence par initialiser les poids de chaque observations à $1/n$ et puis ce poids est ajusté en fonction de la nouvelle estimation, c'est-à-dire, à chaque itération l'algorithme corrige au fur et à mesure l'importance de l'observation en fonction de la qualité de son classement. Le résultat du classifieur boosté est une combinaison des classifieurs $g$ pondérée par les qualités d'ajustement de chaque modèle.

**Algorithme 2.2** Algorithme Adaboost

Initialiser les poids $\omega = \{\omega_i = 1/n \,; i = 1, ..., n\}$ avec n est le nombre d'observations

Pour m=1 à M :

- Estimer le modèle $g_m$ sur l'échantillon d'apprentissage pondéré par les poids $\omega$.
- Calculer le taux d'erreur :

$$\epsilon_m = \frac{\sum_{i=1}^{n} \omega_i \mathbf{1}_{y_i \neq g_m(x_i)}}{\sum_{i=1}^{n} \omega_i}$$

- Calculer $\alpha_m = log((1 - \epsilon_m)/\epsilon_m)$
- Réajuster les pondérations :

$$\omega_i = \omega_i \times exp(\alpha_m \mathbf{1}_{y_i \neq g_m(x_i)}); i = 1, ..., n$$

Fin pour

Résultat du vote : $\hat{f_m}(x) = signe[\sum_{m=1}^{M} \alpha_m g_m(x)]$

Les coefficients $\alpha_m$ sont positifs lorsque les taux d'erreur $\epsilon_m$ sont inférieurs à 0.5. Autrement dit, il faut vérifier que le classifieur est faible mais pas mauvais (Il faut que le taux d'erreur ne soit pas supérieur à 50%).

Cet algorithme repose sur le choix d'un modèle faible g comme classifieur de base (appelé parfois une règle) parmi une famille de modèles $\mathcal{G}$. Dans le cas de la classification binaire, ce choix est fait en minimisant l'espérance de la fonction de perte $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ :

$$g^*(x) = argmin\mathbb{E}[l(Y, g(X))]$$

Naturellement, pour le cas de classification, la fonction de perte est définie comme :

$$l(y, g(x)) = \mathbf{1}_{g(x)\neq y}$$

Comme la loi de (X,Y) est inconnue, l'idée est donc de minimiser l'estimation empirique de $\mathbb{E}[l(Y, g(X))]$

$$g_n^*(x) = \underset{g\in\mathcal{G}}{argmin}\frac{1}{n}\sum_{i=1}^{n} l(Y_i, g(X_i)) = \underset{g\in\mathcal{G}}{argmin}\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i\neq g(X_i)}$$

Ce problème de minimisation n'est pas facile à résoudre pour plusieurs familles de règles. Une solution est adoptée qui consiste à transformer la fonction de perte en une fonction convexe. Dans le cadre de l'algorithme *adaboost*, la fonction de perte correspond à $l(y, g(x)) = exp(-yg(x))$. Ce résultat a été démontré dans [Hastie et al, 2009].

### 2.4.2 Gradient Boosting

Soit $l : \mathbb{R}\times\mathbb{R}$ une fonction convexe. On cherche $g$ permettant de minimiser $\mathbb{E}[l(Y, g(X))]$. Pour trouver le minimum de la fonction de perte convexe, la méthode de descente de gradient est appliquée.

Il est intéressant de rappeler que dans le cas de minimisation d'une fonction strictement convexe $J : \mathbb{R} \to \mathbb{R}$, la méthode de Newton-Raphson est utilisée. Elle consiste à trouver la suite $(x_k)$ qui converge vers la solution du problème de minimisation $\tilde{x}$.

L'initialisation de l'algorithme se fait en choisissant une valeur $x_0$, il suffit ensuite de trouver $x_1 = x_0 + h$ tel que $J'(x_1) \simeq 0$. Un développement limité permet d'obtenir une approximation

$$J'(x_0 + h) \simeq J'(x_0) + hJ"(x_0)$$

Avec la condition initiale, $h = -(J"(x_0))^{-1}J'(x_0)$. Posons $\lambda = (J"(x_0))^{-1}$, donc $x_1 = x_0 - \lambda J'(x_0)$. Par récurrence, $x_k = x_{k-1} - \lambda J'(x_{k-1})$.

Par analogie avec le modèle adaptatif, l'objectif est de minimiser en g la fonction $\mathbb{E}[l(Y, g(X))]$ qui est inconnue. Et ainsi, le problème revient à minimiser la version empirique de la fonction

$$\frac{1}{n}\sum_{i=1}^{n} l(Y_i, g(X_i)$$

[Buhlman & Yu , 2003] et [Hastie et al, 2009] ont transposé l'approche simple au cadre vectoriel et la formule de récurrence s'écrit :

$$g_k(x_i) = g_{k-1}(x_i) - \lambda[\frac{\partial l(y_i, g(x_i))}{\partial g(x_i)}]_{|g(x_i)=g_{k-1}(x_i)}$$

La formule de récurrence permet d'obtenir une suite d'estimateurs $(g_k)_k$ aux points $x_1, \ldots, x_n$. Afin de trouver les estimateurs en tout point $x \in \mathbb{R}^d$, il est nécessaire d'effectuer, à chaque itération, une régression sur l'échantillon $(x_1, U_1), ..., (x_n, U_n)$.

avec

$$U_i = -[\frac{\partial l(y_i, g(x_i))}{\partial g(x_{i)}}]_{|g(x_i)=g_{k-1}(x_i)},\ i = 1, ..., n.$$

Une famille d'algorithmes dont la fonction de perte $l$ est convexe et différentiable a été proposée par Freidman en 2002 sous le nom de GBM *gradient boosting models*. Il s'agit du même principe que l'algorithme adaboost qui consiste à agréger plusieurs modèles en s'approchant de la meilleure solution à chaque itération. Le modèle adaptatif est transformé en une descente de gradient, c'est-à-dire, au lieu de chercher le meilleur classifieur avec adaboost, l'idée consiste à chercher un meilleur pas de descente $\gamma$ à travers l'algorithme suivant :

$$min_\gamma \sum_{i=1}^{n} [l(y_i, g_{m-1}(x_i) - \gamma\frac{\partial l(y_i, g_{m-1}(x_i))}{\partial g_{m-1}(x_i)})].$$

**Algorithme 2.3** Boosting par descente de gradient fonctionnelle

Soit x à prévoir, M le nombre d'itérations et $d_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ l'échantillon d'apprentissage

Initialiser $\hat{g_0} = argmin_\gamma \frac{1}{n} \sum_{i=1}^{n} l(y_i, \gamma)$

Pour m=1 à M :

- Calculer $U_i = -[\frac{\partial l(y_i, g(x_i))}{\partial g(x_i)}]_{|g(x_i)=g_{m-1}(x_i)}$ , i=1,...,n
- Ajuster la règle faible $\delta_m$ (un arbre ) sur l'échantillon $(x_1, U_1), ..., (x_n, U_n)$
- Calculer $\gamma_m$ en résolvant : $min_\gamma \sum_{i=1}^{n} l(y_i, g_{m-1}(x_i) + \gamma\delta_m(x_i))$.
- Mise à jour : $\hat{g}_m(x) = \hat{g}_{m-1}(x) + \gamma_m\delta_m(x)$

Résultat : $\hat{g}_M(x)$

D'autres versions de cet algorithme ont été proposées en incluant un sous-échantillonnage aléatoire à chaque étape afin de construire des *prédicteurs* plus indépendants, ou en ajoutant un coefficient de rétrécissement (*shrinkage*) qui pénalise l'ajout d'un nouveau modèle dans l'agrégation et ralentit la convergence. Dans la pratique, il est nécessaire de contrôler ces paramètres pour un algorithme optimal et éviter le sur-apprentissage.

### 2.4.3 Extreme Gradient Boosting

Il s'agit d'une découverte récente, Chen et Guestrin ont proposé en 2016 l'algorithme d'*extreme gradient boosting*. Cette méthode est une implémentation optimisée de l'algorithme Gradient Boosting. Ce modèle a eu un grand succès et est systématiquement utilisé dans la majorité des solutions des compétitions de machine learning.

La nouveauté dans cet algorithme est l'introduction d'un terme de régularisation $\mathcal{L}$ qui complète la fonction de perte convexe et différentiable.

$$\mathcal{L}(g) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(\delta_m)$$

avec

$$\Omega(\delta) = \alpha|\delta| + \frac{1}{2}\beta \parallel \omega \parallel^2$$

Si le *classifieur* est un arbre de décision, $|\delta|$ correspond au nombre de feuilles de l'arbre $\delta$ et $\omega$ représente le vecteur des valeurs attribuées à chacune des feuilles.

Le terme de régularisation est introduit dans le but d'éviter le sur-apprentissage en limitant l'ajustement de l'arbre à chaque itération.

L'optimisation de l'algorithme nécessite la prise en compte de plusieurs paramètres d'où la complexité de ce modèle. D'autres paramètres s'ajoutent à ceux du gradient *boosting* comme le coefficient de pénalisation $\alpha$ et le coefficient de régularisation $\beta$.

## 2.5 Validation des modèles

En apprentissage supervisé, la construction d'un modèle est faite sur la base d'apprentissage. Ce modèle est ensuite testé sur une base de validation. Pour évaluer le modèle, nous disposons de plusieurs métriques et indicateurs comme la courbe ROC, la matrice de confusion, la déviance , $R^2$, etc.

Nous nous limitons à la matrice de confusion et l'aire sous la courbe ROC dans le cadre de l'évaluation des modèles de classement.

### 2.5.1 Matrice de confusion

L'évaluation d'un modèle de classement binaire (2 classes) peut se faire à l'aide de la matrice de confusion. Il s'agit de confronter les valeurs observées de la variable cible avec les valeurs prédites et de comptabiliser les bonnes et les mauvaises prédictions.

Nous désignons par 1 la classe des frauduleux et par 0 la classe des non frauduleux, la matrice de confusion peut être présentée de la manière suivante :

| | | Observations | | |
|---|---|---|---|---|
| | | 0 | 1 | Total |
| Prédictions | 0 | VN | FN | VN+FN |
| | 1 | FP | VP | FP+VP |
| | Total | VN+FP | FN+VP | VN+FN+FP+VP |

où :

— VN sont les vrais négatifs, autrement dit, les observations qui ont été prédites négatives et qui le sont réellement.

— FN sont les faux négatifs c'est-à-dire les observations qui ont été prédites négatives et qui sont en réalité des positifs.

— FP sont les faux positifs, ils correspondent aux individus qui sont classés positifs alors qu'ils sont des négatifs.

— VP sont les vrais positifs, c'est-à-dire les individus qui sont prédits positifs et qui le sont réellement.

La somme $N = VN + VP + FN + FP$ représente le nombre total des individus.

A partir de la matrice de confusion, nous calculons le taux d'erreur qui est défini comme le rapport entre le nombre des individus issus du mauvais classement et le nombre total de l'effectif :

$$\epsilon = \frac{FN + FP}{N}$$

Cette mesure permet d'estimer le taux de mauvais classement du modèle.

De la même manière, le taux de bon classement (*accuracy* en anglais) est défini comme le complémentaire à 1 du taux d'erreur

$$\theta = \frac{VP + VN}{N} = 1 - \epsilon$$

La sensibilité du modèle (appelé parfois rappel ou *recall* en anglais) permet de mesurer la capacité du modèle à retrouver les positifs (le taux des vrais positifs), elle est définie comme :

$$Sensibilité = \frac{VP}{VP + FN}$$

La spécificité du modèle permet de mesurer la proportion des vrais négatifs parmi tous les négatifs détectés

$$Spécificité = \frac{VN}{VN + FP}$$

La précision (ou bien la valeur prédictive positive VPP) est une mesure de la proportion des vrais positifs parmi tous les individus classés comme positifs

$$Précision = \frac{VP}{VP + FP}$$

Un autre indicateur qui peut être utilisé est la moyenne harmonique ou bien la F-Mesure. Cet indicateur est une combinaison des mesures définies ci dessus

$$F = 2 * \frac{sensibilité * précision}{précision + sensibilité}$$

Ces mesures permettent d'évaluer le modèle et d'estimer s'il s'agit d'un « bon » modèle ou non. Un « bon » modèle correspond à une valeur faible du taux d'erreur (respectivement une *accuracy* proche de 1), des valeurs de précision et de spécificité qui sont proches de 1.

### 2.5.2 Courbe ROC

La courbe ROC (*Receiving Operator Characteristic*) permet d'évaluer et de comparer les performances des estimateurs (*classificateurs*). Graphiquement, la courbe ROC permet de représenter le taux de vrais positifs (la sensibilité) en fonction du taux de faux positifs (1-spécificité) pour différents seuils de classification.

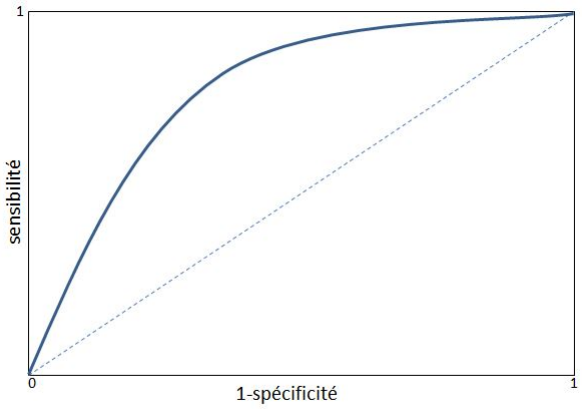



FIGURE 2.1 – La courbe ROC

A partir de la courbe ROC, nous pouvons calculer une métrique numérique en mesurant l'aire sous la courbe, c'est le critère $AUC$. La situation référence correspond à $AUC$ égale à 0.5. Pour augmenter les performances du modèle, il faut que la valeur de *l'*$AUC$ soit supérieure à 0.5.

## 2.6 Techniques de ré-échantillonnage

Dans la majorité des cas, le volume de la base mise à disposition est très important et les deux classes (Fraude ou non) ne sont pas représentées d'une façon équilibrée. Le problème des données déséquilibrées s'explique principalement dans notre cas par la rareté et la difficulté à détecter la fraude.

Pour pallier à ce problème, plusieurs techniques sont envisageables dont le principal enjeu stratégique est de s'assurer une performance de modélisation équivalente à celle du modèle calibré sur la base complète.

Le ré-échantillonnage consiste à construire aléatoirement des sous-échantillons à partir du jeu de données initial. On distingue le *boostrap* et la validation croisée.

### 2.6.1 Bootstrap

Le *boostrap* est une technique de ré-échantillonnage qui consiste à répliquer les données à partir du jeu de données initial. Elle est basée sur la création des nouveaux échantillons par tirage avec remise.

Il s'agit d'une méthode non paramétrique qui permet de générer aléatoirement des échantillons indépendants et identiquement distribués suivant la loi uniforme.

**Algorithme 2.4** Bootstrap

— Fixer le nombre d'échantillons à B

— Répéter pour b allant de 1 à B :

- Tirage avec remise d'un échantillon bootstrap
- Apprentissage du modèle sur l'échantillon bootstrap
- Calcul de l'erreur sur l'échantillon bootstrap

### 2.6.2 Validation croisée

La validation croisée est une application de la technique de ré-échantillonnage pour sélectionner et estimer la fiabilité du modèle. Il existe trois méthodes de validation croisée.

— *Testset validation* ou *holdout*

— *k-fold cross-validation*

— *Leave-one-out cross-validation (LOOCV)*

Soit N la taille de l'échantillon initial.

#### Testset validation

Cette méthode classique de validation croisée consiste à diviser l'échantillon en deux sous-échantillons, le premier représente l'échantillon d'apprentissage (*learning ou training set* en anglais) et comporte de 60% à 80% de

l'échantillon initial, et le second est un échantillon de test ou de validation (*test set* en anglais). Le modèle est construit sur les données de l'échantillon d'apprentissage et évalué sur les données test.

**k-fold cross-validation**

**Algorithme 2.5** k-fold cross-validation

— Diviser l'échantillon initial en K sous-échantillons.

— Répéter pour k allant de 1 à K :

- Sélectionner kième échantillon comme ensemble de validation
- Entrainer le modèle sur (k-1) échantillons
- Valider le modèle avec le kième échantillon
- Estimer l'erreur de prédiction sur l'échantillon de validation $\epsilon_k$

— Calculer l'erreur de prédiction comme la moyenne des k erreurs $\epsilon_k$.

Dans cette méthode, chaque sous-ensemble est utilisé exactement une fois comme un échantillon de validation.

**Leave-one-out cross validation**

Cette méthode est un cas particulier de la méthode *k-fold cross-validation* où K=N. Elle consiste à choisir N-1 observations pour entrainer le modèle et le valider sur la énième observation. Cette opération est répétée N fois afin que toutes les observations soient utilisées comme un ensemble de validation.

La validation croisée permet d'éviter le problème de sur-apprentissage [7]. Le risque de sur-apprentissage désigne l'adaptation excessive du modèle aux données. Le modèle capturera non seulement toutes les interactions entre données mais aussi le bruit produit par ces données. Ce problème se caractérise par une variance très élevée.

Le choix de la méthode de validation croisée est basé sur un compromis entre le temps de calcul et la performance du modèle. Le découpage apprentissage / test ou validation est intéressant sur les grandes bases. Les

7. En contre partie, le risque de sous-apprentissage consiste à une mauvaise adaptation du modèle aux données et le modèle n'arrive pas à capturer les interactions entre ces données. Le modèle a ainsi un pouvoir prédictif faible et souffre d'un grand biais.

techniques *k-fold* ou *leave-one-out* sont utilisées généralement dans le tuning du modèle, c'est-à-dire dans l'optimisation des hyper-paramètres.

### 2.6.3 Resampling

Dans la détection de la fraude, la classe des positifs représente environ 1% du jeu de données. Construire un modèle sur une telle base induit le risque d'avoir un modèle peu performant et qui prédit toujours la classe majoritaire. Or, dans notre cas, nous nous intéressons à la classe minoritaire. L'échantillonnage est introduit pour gérer une répartition déséquilibrée des classes dans la base. Cette technique consiste à augmenter le taux des positifs ou à diminuer le taux des négatifs dans la base.

Il existe trois méthodes d'échantillonnage visant à rééquilibrer et redresser la distribution asymétrique des classes :

#### OverSampling ou UpSampling : Sur-échantillonnage

Cette méthode consiste à augmenter le nombre des individus de la classe minoritaire (les positifs). Cependant, cette méthode augmente le risque du sur-ajustement.

#### UnderSampling ou DownSampling : Sous-échantillonnage

C'est une approche simple à mettre en œuvre, il s'agit de supprimer d'une façon aléatoire les individus de la classe majoritaire (les négatifs). En revanche, cette méthode intègre le risque de supprimer des observations représentatives dans la base.

#### Smote : Synthetic Minority Over-Sampling Technique

Cette technique se caractérise par la génération artificielle des individus synthétiques. En effet, l'algorithme cherche les k plus proches voisins des individus de la classe minoritaire puis il synthétise aléatoirement de nouveaux individus entre ces deux points selon le taux de sur-échantillonnage voulu.

Chapitre 3

# Les données

## 3.1 Présentation des bases de données

Notre étude s'appuiera sur deux bases de données principales :

— Une base consolidée des sinistres de la marque *MAAF*, survenus entre 2012 et 2018.

— Une base *Stat_Fraude* qui est le produit des pistes de fraudes identifiées par les équipes Sinistres en charge de la lutte anti-fraude.

**Base Sinistre** : Les données sont constituées des éléments relatifs aux circonstances du sinistre, des informations sur le contrat et le client.

Nous ciblerons les sinistres produits par des personnes physiques, qui désignent les clients individuels : les particuliers (les personnes morales correspondent à des entités juridiques, c'est-à-dire les entreprises).

La base *Sinistre* finale est le résultat de la fusion de différentes bases dont nous disposons :

— Une base des sinistres clôturés et en cours. Chaque ligne de la base correspond à un dossier indiquant ainsi le dernier état du sinistre. L'identifiant de chaque observation est le numéro du sinistre. Cette base comporte 32 colonnes correspondant à des informations sur les circonstances du sinistre telles que la date et le lieu de survenance, le type du sinistre, le contrat et la garantie mise en jeu, etc.

— Une base des clients contenant les informations sur les assurés (les personnes physiques) comme la date de naissance, l'adresse postale, la catégorie professionnelle, etc...

— Une base des contrats qui contient l'historique des contrats automobile et les éventuelles modifications apportées sur la police

— Les données *DARVA*, il s'agit des rapports d'expertise, les points de choc, des données sur les réparateurs et les factures et les données *FIDELIA* qui est la compagnie en charge de l'indemnisation des garanties assistances au sein de *Covéa*.

**Base *STAT_FRAUDE :*** ce *DataSet* contient les sinistres détectés comme potentiellement frauduleux. En effet, nous y trouvons tous les dossiers sinistres faisant l'objet d'une analyse lutte anti-fraude dès que les premiers soupçons de fraude sont confirmés et qu'il apparait que des investigations complémentaires sont nécessaires. Cette base contient également des informations relatives à cette détection comme :

— la date de la détection

— l'origine de cette détection c'est-à-dire l'entité responsable de cette détection. Nous distinguons alors les gestionnaires de sinistres ou bien le prestataire externe [1].

— la typologie de fraude. Il s'agit d'une catégorisation des sinistres suspectés comme frauduleux. Par exemple, nous pouvons trouver comme scénario *sinistre monté* ou bien *dommages antérieurs*...

— le coût qui est le montant de l'indemnité impacté par le risque de fraude et le gain qui est le montant qui a été gagné au titre de la fraude, il s'agit de l'économie brute sans déduction d'éventuels frais d'enquête, d'analyses, de justice ou de charges salariales.

Une fois la base est construite, la préparation, le nettoyage et le traitement des données représentent une phase préliminaire d'un projet en *Machine Learning*. Cette étape est appelée souvent *data preprocessing* ou bien *data preparation*.

Dans cette partie, nous allons présenter le processus de préparation de la base ainsi que la création et le traitement des variables pour entamer la modélisation de la fraude.

### 3.1.1 Nettoyage de la base et traitement des données manquantes

La basse contient initialement plus de 100 variables. Ces données peuvent être des variables quantitatives (généralement au format numérique comme

1. Covéa confie la tâche de la détection automatisée de la fraude en auto à un partenaire externe.

la valeur du véhicule), des variables qualitatives (nominales comme la situation maritale) et des variables temporelles au format date.

Le processus du nettoyage de la base est effectué en plusieurs étapes :

Premièrement, certaines variables ont été éliminées car elles apportent la même information et car elles sont généralement fortement corrélées, par exemple l'âge et l'ancienneté de permis. De plus, les variables qui ne sont pas utilisées dans la modélisation comme le numéro de contrat, le nom et le prénom ont été également supprimées de la base.

Les variables au format date ne peuvent pas être directement utilisées dans la modélisation, mais il est judicieux de récupérer autrement les informations qu'elles rapportent. En effet, certaines dates présentes dans la base sont transformées en variables quantitatives ou qualitatives, par exemple :

- La date de naissance est transformée en variable âge
- L'écart entre la date de survenance et la date de déclaration du sinistre est évalué
- Le jour du survenance du sinistre a été récupéré de la date de survenance.

Par ailleurs, un premier modèle de classification (*Xgboost*) a été testé sur la base initiale afin d'avoir une idée sur les variables les plus importantes et qui contribuent dans l'explication de la fraude et d'effectuer ainsi une première sélection.

Ce processus a permis de réduire significativement la taille de la base et d'avoir une idée sur la qualité des données.

La majorité des données est disponible et complète dans la base présélectionnée. Vu le pourcentage faible des données manquantes, la stratégie de complétion adoptée est simple, soit une imputation par la moyenne ou la médiane pour les données quantitatives (par exemple, une valeur à neuf manquante est remplacée par la médiane), soit une imputation par la classe la plus présente c'est-à-dire la plus observée pour les variables qualitatives. En revanche, une modalité non renseignée peut avoir une interprétation particulière. Dans ce cas, il est intéressant de créer une nouvelle modalité pour ces variables qualitatives.

### 3.1.2 Création de nouveaux indicateurs

Dans ce projet, la fraude est considérée comme une information catégorielle. Nous distinguons deux classes, les cas de fraude et les cas de non fraude à l'aide d'une variable binaire.

### La variable Fraude

Nous construisons la variable *Fraude* à partir d'une base contenant la liste des sinistres détectés comme frauduleux (suspectées et avérées). Il s'agit d'une variable binaire qui prend la valeur 0 pour désigner un sinistre non frauduleux et 1 pour un sinistre frauduleux. Notre étude repose sur une hypothèse forte, nous supposons que l'ensemble des sinistres labellisés comme non frauduleux dans le passé étaient effectivement non frauduleux.

La liste des sinistres frauduleux contient les fraudes détectées par les gestionnaires d'une part et des sinistres remontés par le prestataire externe d'autre part. Dans cette base, nous distinguons la fraude avérée et la fraude non avérée. La notion de fraude avérée est un indicateur précisant si la fraude est établie. Selon l'agence *ALFA*, elle est établie dès lors que l'assureur peut justifier le fait qui produit l'extinction de l'obligation d'indemnisation (nullité, exclusion, déchéance) ou apporter la preuve d'une situation de non assurance (dissimulée volontairement).

### Les scénarios métier

Nous disposons d'une liste de scénarios appelés souvent *les règles métier* qui permettent aux gestionnaires de remonter des alertes de suspicion afin de détecter les comportements frauduleux. ces *règles métier* ont été produites empiriquement sur l'expertise des conseillers spécialisés en recherche de fraude sur la base des dossiers investigués. Un scénario de fraude est constitué d'une combinaison d'indicateurs ou des règles. Un indicateur correspond généralement à une donnée liée aux circonstances du sinistre, à des modifications intervenues au contrat ou bien à des calculs techniques éventuellement utiles pour réaliser des constations permettant de mettre en évidence une fraude potentielle.

A partir de cette liste, nous avons créé de nouvelles variables permettant ainsi d'une part, d'enrichir la base et d'avoir d'autres indicateurs explicatifs de la fraude autre que les informations générales sur les sinistres. D'autre part, ces indicateurs permettent également de faciliter la tâche d'investigation pour les gestionnaires en affectant le sinistre frauduleux à une typologie de fraude, par exemple, la souscription récente d'un contrat.

Les variables que nous avons ajoutées sont :

— **Souscription récente** : Il s'agit de calculer le décalage entre la date de souscription et la date de survenance du sinistre. Une variable binaire est créée si ce décalage est inférieur à un délai que nous avons

défini de 6 mois. Ce seuil est fixé pour toutes les variables suivantes.

- **Avenant récent** : Il s'agit de voir s'il y a eu une modification du contrat notamment l'ajout d'une garantie, et de calculer le décalage entre la date de l'ajout de la garantie et la date du sinistre. Une variable binaire est créée si ce décalage est inférieur au seuil fixé.
- **Sinistre récent :** Cette variable binaire consiste à évaluer le délai entre deux sinistres de chaque assuré. Un seuil à 6 mois a été positionné en deçà duquel le sinistre est considéré comme récent.
- **Assistance récente :** Elle correspond à une variable binaire pour caractériser la présence d'un dossier d'assistance récent (moins de 6 mois) pour une panne ou un accident.
- **Nombre de pannes**
- **Nombre de pannes non prises en charge**
- Un ratio est calculé à partir du nombre de pannes non prises en charge par rapport au nombre total de pannes pour chaque assuré.
- **Nombre de sinistres**
- **Nombre de contrats du client**
- Un ratio est évalué à partir du nombre de sinistres par rapport au nombre de contrats.
- Décalage entre la date d'ouverture du dossier sinistre et la date de survenance du sinistre.
- Une variable binaire permet de vérifier si le jour de la survenance du sinistre correspond à un jour férié.

### Utilisation des données externes

Nous disposons des informations sur le lieu de survenance de sinistre, principalement le code de département. Ceci a permis de créer la variable *Région* . Il existe 18 régions administratives définies par des codes INSEE.

Notre objectif est de réduire le nombre de modalités de l'information concernant le lieu du sinistre. En effet, au lieu d'avoir une variable comportant environ 101 modalités (le code de département), nous avons ainsi une variable avec 18 modalités (le nombre de régions).

Nous avons également utilisé une base publiée par le ministère de l'Intérieur concernant la délinquance en France en 2016. Cette base contient le taux de vol de véhicules dans chaque département. Ces statistiques ont servi à

créer une variable qui permet de classer les départements en 3 classes : département avec un faible taux de vol, département avec un taux de vol moyen et département avec un taux de vol élevé. Nous avons attribué respectivement à ces 3 classes les modalités 0,1 et 2.

## 3.2 Analyse descriptive de la base

Nous possédons un portefeuille de sinistres survenus sur 7 ans d'exercice. Nous nous intéressons dans ce travail aux sinistres matériels [2] hors les sinistres Bris de glace, les sinistres avec un tiers non identifié et les sinistres avec un animal.

Dans cette partie, nous allons présenter la répartition des données dans la base, la relation et la corrélation des variables explicatives entre elles d'une part, et avec la variable cible *Fraude* d'autre part.

### 3.2.1 Statistiques descriptives

#### Année de survenance du sinistre

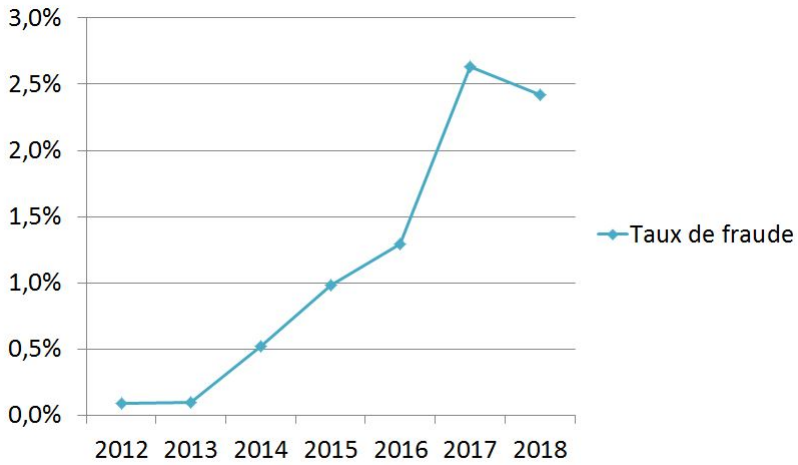



FIGURE 3.1 – Le taux de fraude dans la base par année de survenance du sinistre

2. Les sinistres corporels sont supprimés de la base et font l'objet d'une autre étude de détection de fraude.

Rappelons que nous disposons d'un historique de 7 ans des sinistres automobile. Nous avons calculé le pourcentage de fraude dans la base de tous les sinistres en fonction de l'année de survenance. Ce taux est issu du référentiel *StatFraude*. En moyenne, le taux de fraude est de l'ordre de 1% et le taux le plus élevé est atteint en 2017. Pour les années de 2012 à 2016, le nombre de sinistres frauduleux est très faible, ce qui se traduit par un taux de fraude inférieur à 0.5% en moyenne. Ceci est expliqué par le fait que la base *StatFraude* n'a été alimentée d'une manière exhaustive que depuis 2017. L'hypothèse fondamentale de notre étude (un sinistre historique non frauduleux était bien non frauduleux) est très fragile pour ces années. Nous avons donc intérêt à ne pas prendre un historique d'étude trop long, car les années les plus anciennes sont les plus biaisées.

Alors, nous avons choisi de réduire la base des sinistres en travaillant sur un historique de 2 ans, soit les deux années 2017 et 2018.

### Évènement du sinistre

Dans un second temps, nous avons évalué le taux de fraude en fonction de l'évènement du sinistre autrement dit la principale garantie mise en jeu lors du sinistre. Nous distinguons alors *circulation*, *incendie*, *vol*, *force de la nature*, *bris de glace* et *autre*.

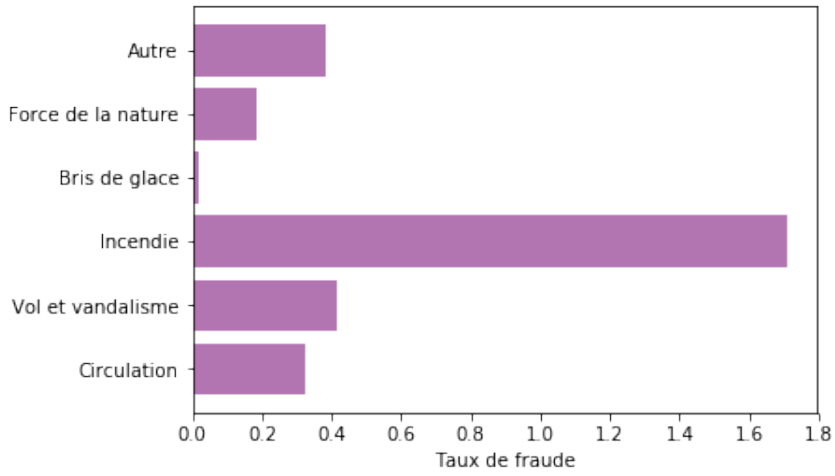



FIGURE 3.2 – Taux de fraude par type d'évènement du sinistre

Le taux de fraude correspond au rapport entre le nombre de cas de fraude et le nombre total de sinistres de chacun des évènements.

A partir de ce graphique et comme évoqué précédemment, nous nous intéressons aux trois grandes catégories de sinistres, notamment la circulation, l'incendie et le vol (Les sinistres vol, vandalisme et tentative de vol sont regroupés). Nous supprimons ainsi les autres sinistres de la base ce qui permet de réduire la taille de la base initiale.

Ce choix est justifié également du fait que certains indicateurs sont spécifiques à l'évènement du sinistre et à la garantie mise en jeu. Par exemple, pour les sinistres *vol*, l'indicateur *véhicule retrouvé après le vol,* permettant de préciser si le véhicule a été retrouvé après la déclaration du vol, est propre à cette catégorie de sinistres.

Les garanties incendie, vol et vandalisme sont particulièrement sensibles au risque de fraude car le coût moyen des indemnisations est plus élevé pour ces garanties. De plus, il est particulièrement difficile pour les experts de prouver qu'un incendie volontaire a été réalisé ou orchestré par l'assuré ainsi que le vol total notamment lorsque le véhicule n'est pas retrouvé.

**Ancienneté de souscription du contrat**

L'écart entre la date de souscription du contrat et la date de survenance du sinistre est transformé en une variable qualitative ayant 6 modalités. Pour chaque classe, le taux de fraude est évalué en divisant le nombre de sinistre frauduleux par le nombre total des sinistres dans cette classe.


FIGURE 3.3 – Taux de fraude par ancienneté du contrat

Par souci de confidentialité, les valeurs de la variable *nombre de sinistre* ne sont pas affichées dans le graphique.

Les différentes modalités de la variables ancienneté du contrat sont obtenues à l'aide de la méthode des quantiles et correspondent à :

- Classe 0 : l'écart entre la date de souscription et la date de sinistre est inférieur à 1 an.
- Classe 1 : l'écart entre la date de souscription et la date de sinistre est entre 1 et 3 ans.
- Classe 2 : l'écart entre la date de souscription et la date de sinistre est entre 3 et 6 ans.
- Classe 3 : l'écart entre la date de souscription et la date de sinistre est entre 6 et 10 ans.
- Classe 4 : l'écart entre la date de souscription et la date de sinistre est entre 10 et 18 ans.
- Classe 5 : l'écart entre la date de souscription et la date de sinistre est supérieur à 18 ans.

Le graphique ci-dessus montre que le taux de fraude pour les contrats souscrits au plus tard 1 an avant le sinistre est le plus élevé. Plus le contrat est vieux, plus le taux de fraude est faible. Ceci confirme aussi la pertinence de l'indicateur métier *souscription récente* proposé par les gestionnaires du

pôle lutte contre la fraude. En effet, certains risques de fraude sont plus présents dans la première année qui suit la souscription du contrat notamment la fraude opportuniste. Il s'agit de déclarer un sinistre suite à l'achat d'un véhicule épave et de profiter ainsi de l'indemnisation pour le réparer.

**L'âge de l'assuré**

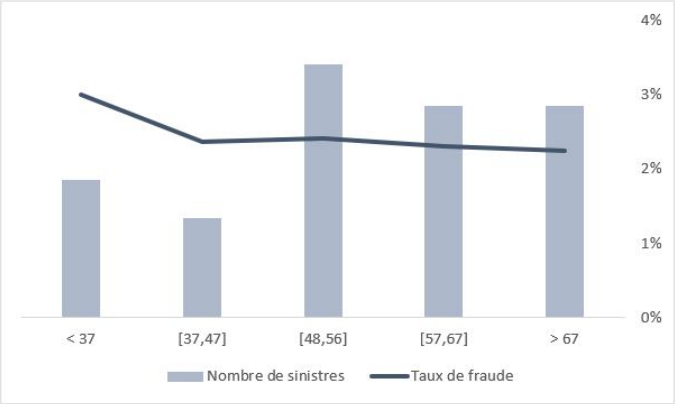



FIGURE 3.4 – Taux de fraude par tranche d'âge

A l'aide de la méthode de quantile, nous avons créé cinq tranches d'âge. L'évaluation du taux de fraude pour chaque classe montre que ce taux est plus élevé pour la tranche d'âge inférieur à 37 ans. Les jeunes sont considérés comme des profils « à risque » et sont donc les moins bien lotis en matière de prime d'assurance. Ce profils ont tendance à faire des tentatives de fraude plus que les autres tranches d'âges.

**Le jour de la semaine du sinistre**

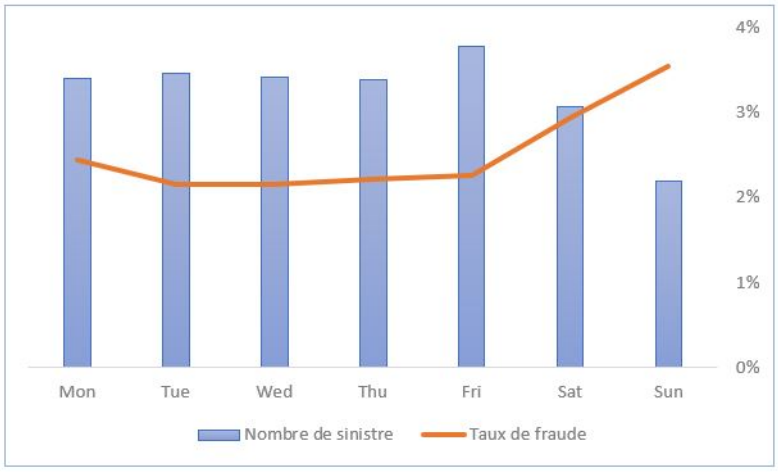



FIGURE 3.5 – Taux de fraude par jour du sinistre

Afin de mieux explorer les données et en se basant sur les connaissances des gestionnaires de sinistres frauduleux, la variable *jour de sinistre* a été créée à partir de la date de survenance du sinistre. Le taux de fraude augmente à partir du vendredi et atteint son maximum le dimanche.

Il s'agit probablement des cas de fraude portant sur les circonstances du sinistre. Le risque ici est de dissimuler les vraies conditions de réalisation du sinistre notamment la modification de l'identité réelle du conducteur en raison de la présence des stupéfiant, de son état alcoolique, de l'absence de son permis de conduire en cours de validité.

**Autres constatations :**

— 99% des sinistres frauduleux n'ont pas un rapport de police (procès verbal PV).

— 30% des sinistres frauduleux sont survenus entre samedi et dimanche.

— 23% des sinistres frauduleux succèdent un autre sinistre déclaré.

### 3.2.2 Étude des corrélations et sélection des variables

L'évaluation de la corrélation des variables est nécessaire avant l'application de tout modèle. Le choix de test d'indépendance dépend du type des variables. Pour les variables catégorielles, nous utilisons les test de Khi-deux ou bien la V de Cramer pour mesurer leur corrélation. Nous pouvons également évaluer les associations entre les modalités de chaque variable. L'évaluation de la corrélation entre deux variables numériques est réalisée en utilisant le test de *Pearson*[3].

#### Test d'indépendance de khi-deux

Le test de *Khi-deux* permet de tester l'indépendance entre deux variables. Sous l'hypothèse nulle, la statistique de ce test suit une loi du *Khi-deux* $\chi^2$.

$H_0$ : Les deux variables sont indépendantes.

$H_1$ : Il existe un lien statistique entre les deux variables.

Pour un seuil de significativité de 5%, l'hypothèse nulle est rejetée lorsque la p-valeur est inférieure au seuil fixé.

#### Test de V de Cramer

Le test de Khi-deux permet de savoir s'il y a un lien entre deux variables et de mesurer la corrélation. La mesure *V de Cramer* permet d'évaluer l'intensité de cette association. Elle correspond à

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \times min(l-1, c-1)}}$$

où n est le nombre d'effectif, l est le nombre de lignes du tableau de contingence[4] et c est le nombre de colonnes de ce tableau.

Plus la valeur du V de Cramer est proche de zéro, plus les deux variables sont indépendantes. Si le V de Cramer est égal à 1 alors les deux variables sont complètement dépendantes.

Pour étudier les corrélations entre deux variables catégorielles, le test de Khi-deux et le test de V de Cramer sont utilisés.

3. Voir Annexe 4.

4. Le tableau de contingence croisant les couples de variables indique les co-occurrences entre les différentes valeurs des variables.

### Analyse de la variance

Pour analyser les corrélations entre une variable catégorielle et une variable continue (numérique), le modèle statistique Analyse de la variance *Anova* peut être utilisé. L'analyse de la variance peut être vue comme une généralisation du test de *Student*. Le test *Anova* est appliqué afin de mesurer l'influence des modalités d'une variable catégorielle sur la loi d'une variable continue. Ce test repose sur le calcul des variances inter-classes et les variances intra-classes, et consiste à rejeter l'égalité des moyennes des échantillons. Cette méthode repose sur la normalité [5] des distributions et l'indépendance des échantillons sous l'hypothèse nulle.

Il suffit de comparer la p-valeur à un seuil de significativité de 5% pour rejeter l'hypothèse $H_0$.

## Analyse des variables

Dans cette partie, nous nous limitons aux variables les plus importantes dans la construction des modèles.

### La valeur du véhicule

Dans notre base de données, la valeur du véhicule correspond à la valeur à neuf du véhicule. C'est la valeur catalogue du véhicule acheté neuf. Cette donnée est importante car certaines garanties permettent d'indemniser le client en fonction de la valeur à neuf du véhicule, sous certaines conditions. Les experts observent généralement un taux de fraude important les semaines ou les mois qui précèdent la fin de l'option valeur à neuf.

Le test Anova a été appliqué pour mesurer l'indépendance entre la variable cible, la fraude, et cette variable. En comparant la p-valeur au seuil de 5%, nous pouvons conclure que la valeur à neuf a une influence sur la variable cible. La moyenne des valeurs de véhicules des sinistres frauduleux est de l'ordre de 22000 €, tandis que pour les sinistres frauduleux, la valeur moyenne des véhicules est de l'ordre de 28000 €.

5. L'hypothèse de normalité n'est pas toujours vérifiée, ni vérifiable en pratique. Pour les échantillons de grande taille, le théorème central limite assure la normalité asymptotique des moyennes empiriques.

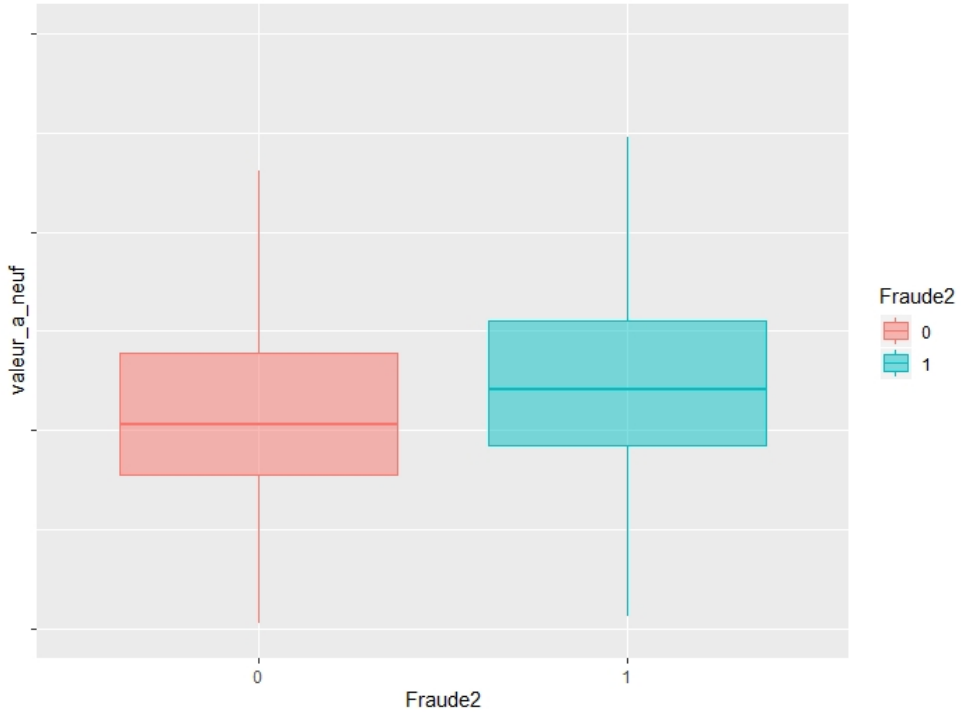



FIGURE 3.6 – Boxplot de la variable valeur à neuf en fonction de la fraude

### Corrélation entre les variables catégorielles

Le test de Khi-deux permet d'identifier l'existence d'une relation entre deux variables catégorielles et de rejeter l'hypothèse de l'indépendance. Par ailleurs, le test de Cramer permet de mesurer l'intensité de cette association si elle existe. Dans le graphique suivant, les deux tests ont été évalués pour les six variables [6]. La taille du cercle permet de renseigner sur l'intensité de la relation entre les deux variables. Plus le cercle est grand, plus la corrélation est forte. Si les deux variables sont indépendantes, le cercle correspond à un point, tel est le cas pour les deux variables *LB_jour_sinistre* et *qualification payeur*.

6. Nous nous limitons à cette liste de variable pour assurer la lisibilité du graphique

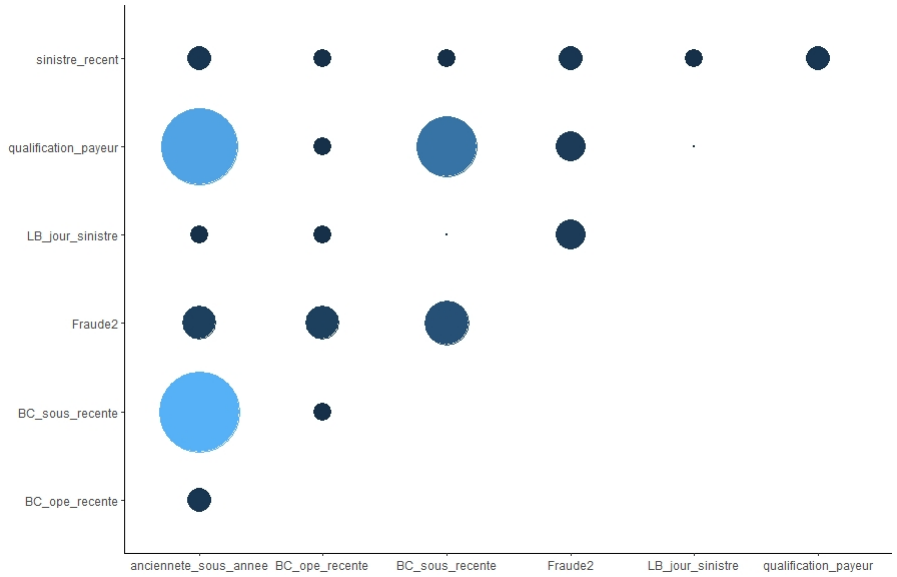



FIGURE 3.7 – Corrélation entre les variables catégorielles

Les deux variables *BC_sous_recente* et *anciennete_sous_annee* sont fortement corrélées car l'indicateur binaire *BC_sous_recente* correspond aux contrats dont l'ancienneté est inférieure à 6 mois. La variable *qualification payeur* renseigne sur la qualité du paiement du souscripteur. Cette variable présente un pouvoir discriminant dans la modélisation de la fraude.

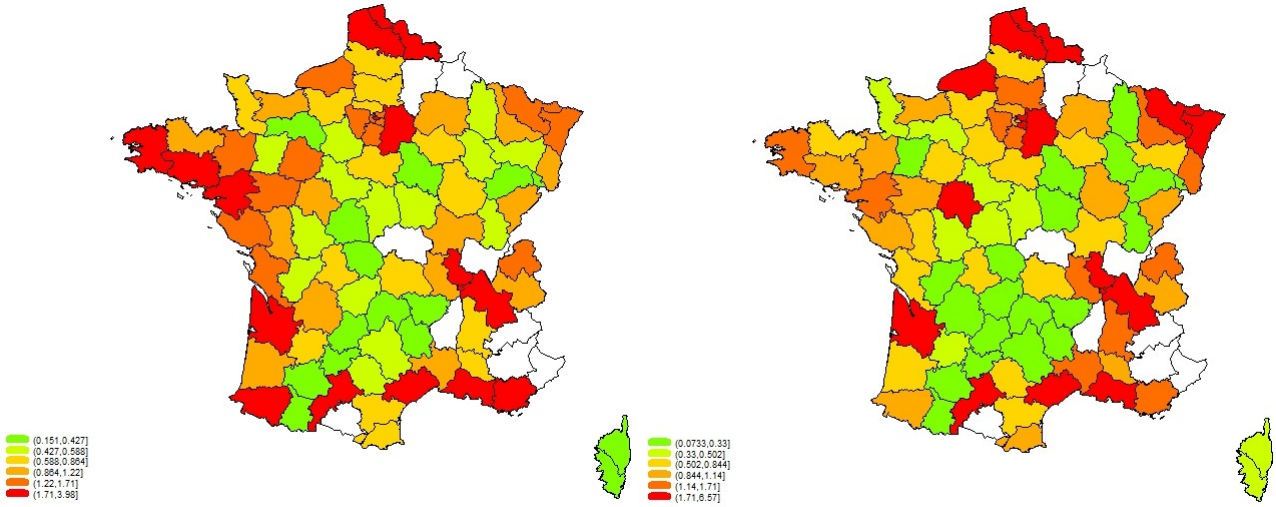
L'analyse de la corrélation entre la variable réponse et les variables explicatives confirme également certaines hypothèses, par exemple l'existence d'une forte relation entre la fraude et une souscription récente du contrat, un lien fort entre les difficultés financières de l'assuré, autrement la difficulté à payer ses cotisations et la tentation de frauder.

### 3.2.3 Retraitement des variables

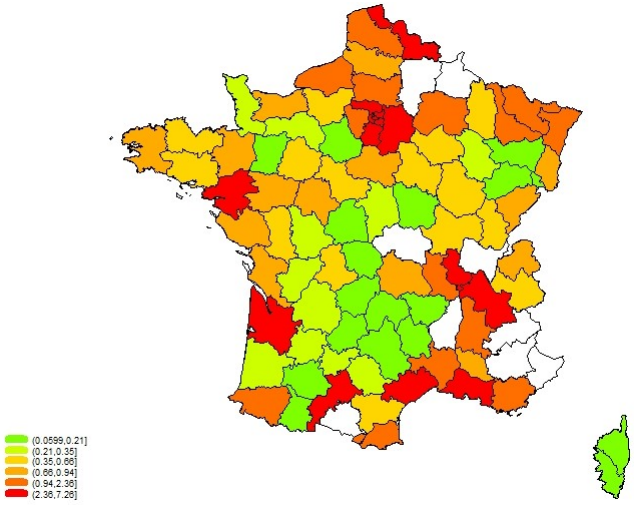
#### 3.2.3.1 Retraitement du code de département

Dans la base des données sinistre, nous disposons des informations liées au lieu de survenance du sinistre. Parmi ces informations, nous trouvons le code de département. Cette variable possède environ 100 modalités. Comme évoqué précédemment, nous avons transformé cette variable en une variable région pour réduire le nombre de modalités. En outre, nous nous sommes

intéressés à la disparité géographique de la fraude. En effet, nous avons défini un ratio : le nombre de sinistres frauduleux par département par rapport au nombre total des sinistres frauduleux. Pour chaque base, nous avons présenté ce ratio sous forme de la carte de France.




Circulation Incendie




Vol

FIGURE 3.8 – Taux de fraude par département

Nous remarquons que ce ratio est élevé dans certaines régions de la France. Ce résultat peut être généralisé pour les trois évènements de sinistre c'est-à-dire pour les trois bases. Pour garder cette information et rappelons que nous avons opté à la transformation de cette variable, nous avons créé un autre indicateur qui permet d'affecter chaque département à une classe. En effet, en utilisant ce ratio, nous avons défini une variable catégorielle de trois classes : la classe 0 pour désigner les départements avec un taux de fraude faible, la classe 1 correspond aux départements dont le ratio est moyen et la classe 2 pour représenter les départements avec un taux de fraude élevé. Pour ce faire, nous avons fait appel à la méthode des quantiles qui permet la discrétisation d'une variable quantitative et la transformer en une variable qualitative.

#### 3.2.3.2 Retraitement de la marque du véhicule

Une analyse de la répartition de la fraude par marque a permis de produire une liste des marques présentant un nombre de cas de fraude élevé. Certes, cette répartition dépend du nombre de sinistres survenus pour chaque marque.

Le graphique ci-dessous montre que les trois marques *Renault*, *Peugeot* et *Citroen* sont les plus présentes dans la base en terme de nombre de fraude.

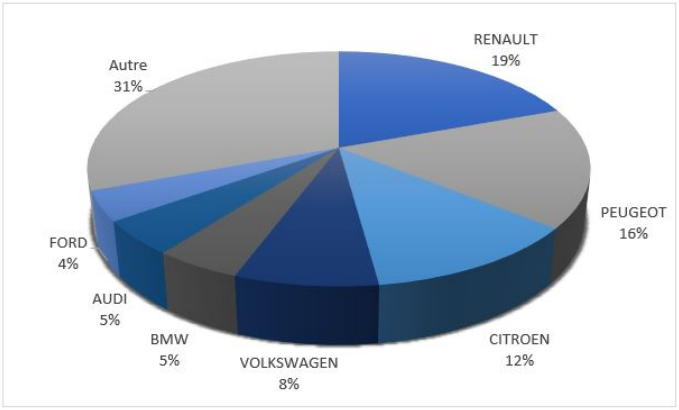



FIGURE 3.9 – Répartition du nombre de fraude par marque

Ce résultat se généralise sur les trois bases. Nous trouvons quasiment la même répartition pour les trois évènements de sinistre. Cependant, afin d'exploiter cette variable, il es nécessaire de la retraiter.

La variable *marque* possède plusieurs modalités. Nous avons autant de modalités que de marques dans la base. Nous avons donc décidé de créer des nouvelles variables en évaluant le taux de fraude par marque. En effet, le premier indicateur consiste à évaluer le ratio : nombre de sinistres frauduleux par rapport au nombre total de sinistre pour chaque marque. Une variable binaire est ainsi créée sous cette condition : si le score est supérieur à la moyenne, la valeur 1 est attribuée à la variable, sinon la variable prend la valeur 0. Le deuxième indicateur consiste à calculer le score : nombre de sinistres frauduleux par marque par rapport au nombre total des cas de fraude dans la base. Et ainsi, une variable binaire a été ajoutée de la même manière que la première variable afin de classer les marques. Il en découle que pour chaque marque, nous avons deux scores qui ont été transformés en deux variables binaires.

Cette procédure permet de faciliter l'entrainement des modèles et de diminuer le temps de calcul. En effet, certains modèles nécessitent la transformation des variables catégorielles en variables *dummies*, c'est-à-dire des variables binaires pour chaque modalité et par conséquent, nous avons autant de variables que de modalités.

Chapitre 4

# Modélisation

## 4.1 Approche globale de la modélisation

Dans ce chapitre, nous allons présenter la calibration des algorithmes d'apprentissage supervisé dans le cadre de la modélisation de la fraude. Les méthodes étudiées sont la régression logistique, les forêts aléatoires et *Extreme Gradient Boosting*.

Dans la première partie de ce chapitre, nous allons rappeler le périmètre de notre recherche ainsi que l'approche adoptée dans cette étude. La deuxième partie de ce chapitre concerne l'application des modèles et les démarches suivies afin d'améliorer leurs résultats.

### 4.1.1 Périmètre de la recherche

Comme nous l'avons évoqué précédemment, nous avons choisi de diviser la base initiale en trois parties en fonction de l'évènement du sinistre. Pour modéliser la fraude, nous avons appliqué les mêmes modèles sur les trois bases. La performance des modèles varie en fonction des variables introduites dans la base étudiée. Certaines variables sont spécifiques à la garantie mise en jeu lors du sinistre et donc elles ne peuvent figurer que dans une seule base.

Rappelons que les cas de fraude confirmés sont faibles par rapport aux cas non frauduleux. De ce fait, la classe des positifs (cas frauduleux) est sous représentée dans la base.

Pour gérer ce problème, la stratégie adoptée est d'équilibrer artificiellement les données. Nous optons à la comparaison de deux méthodes de

ré-échantillonnage, *downSampling* et *Smote*.

### 4.1.2 Application des modèles

Cette partie est consacrée à la calibration des modèles choisis sur les trois bases. La première étape consiste à diviser chacune des trois bases en deux échantillons. Le premier constitue un échantillon d'apprentissage qui sert à entrainer le modèle, et le deuxième permet de le valider. Un échantillonnage aléatoire a été effectué en affectant 80% des observations à la base d'apprentissage et le reste à la base de validation.

Il existe plusieurs critères permettant d'évaluer la qualité d'ajustement du modèle. Dans le cadre de la classification binaire, l'évaluation de la performance du modèle peut être mesurée à l'aide de la matrice de confusion en utilisant le taux de bon classement, la précision et la sensibilité ou bien à l'aide de la courbe ROC. Afin de comparer les modèles sélectionnés, il est judicieux de choisir une mesure de performance valable pour tous les modèles.

Par ailleurs, pour évaluer les écarts entre les valeurs réelles et les valeurs prédites par le modèle, la principale mesure du pouvoir prédictif choisie est l'aire sous la courbe ROC $AUC$. Le modèle est entrainé de façon à maximiser la valeur de l'$AUC$. Par la suite, les prédictions sont calculées en appliquant le modèle sur la base de validation.

Nous avons choisi de ré-échantillonner les bases en appliquant la méthode de sous-échantillonnage *Down Sampling*. Rappelons que cette méthode consiste à diminuer le taux des négatifs dans l'échantillon pour avoir des proportions équilibrées des classes de positifs et de négatifs dans la base. Autrement dit, dans la base d'apprentissage, nous diminuons le nombre des cas non frauduleux afin d'améliorer la sensibilité du modèle et sa capacité à retrouver les cas frauduleux. Ensuite, nous avons recours à la validation croisée pour améliorer les performances du modèle. La méthode de validation croisée appliquée est *K-Fold cross validation*.

## 4.2 La régression logistique

C'est la méthode la plus utilisée en assurance. Il s'agit d'un cas particulier du modèle linéaire généralisé. La régression logistique est un modèle non paramétrique et qui a souvent montré de bonnes performances dans les problèmes de classification.

Les bases de données initiales *circulation*, *vol* et *incendie* contiennent respectivement 33, 35 et 30 variables.

Lorsque la régression logistique est calibrée sur la base, les modalités des variables catégorielles sont transformées en variables binaires. Donc, il y a autant de coefficients à estimer que de modalités pour ce modèle.

Pour évaluer la performance du modèle, il suffit de comparer les valeurs observées et les valeurs prédites de la variable à expliquer. Cependant, le modèle ne prédit pas exactement une classe mais plutôt une probabilité ou un score compris entre 0 et 1. Afin d'attribuer une classe au score obtenu, ce dernier est comparé à un seuil. Si la prédiction est supérieur à ce seuil, la classe « positif » est attribuée, dans le cas contraire, la classe « négatif » est attribuée. Dans notre cas, la classe « positif » correspond à l'évènement fraude. Par défaut, le seuil est fixé à 0.5. Or, ce seuil peut être différent en fonction des données. Pour choisir le seuil optimal, les mesures spécificité et sensibilité sont utilisées. Le seuil est choisi de façon à avoir des niveaux de spécificité et de sensibilité proches et élevés.

### Sélection des variables

La procédure de nettoyage de la base initiale et l'étude de corrélation entre les *prédicteurs* a permis de réduire le nombre de variables explicatives à environ 30 variables dans chaque base. cette étape permet de restreindre le nombre de variables significativement discriminantes et de diminuer la complexité du modèle.

En utilisant le test de WALD [1], les variables introduites dans le modèle ne sont pas toutes pertinentes pour un seuil de significativité égal à 5% (en comparant la p-valeur [2] à un niveau de significativité égal à 5%). Pour simplifier le modèle tout en conservant ses qualités prédictives, nous avons eu recours aux méthodes de sélection des variables.

Plusieurs méthodes sont disponibles pour réduire le nombre de variables dans le modèle. La méthode ascendante ou *Forward* consiste à introduire les variables une par une en commençant par la plus significative. La procédure s'arrête lorsque toutes les variables dans le modèle sont pertinentes. La méthode descendante ou *Backward* commence par considérer un modèle complet, c'est-à-dire un modèle qui inclut toutes les variables explicatives, puis elle procède à l'élimination progressive des variables les moins significatives. La condition d'arrêt est la même que dans la méthode *Forward*. La troisième méthode est une méthode mixte *Stepwise*. Il s'agit d'une combinai-

1. Voir annexe 1.
2. Sous l'hypothèse nulle, la p-valeur correspond à la probabilité d'obtenir une statistique plus grande que la valeur observée. Pour un seuil de significativité $\alpha$, si la p-valeur est inférieure à $\alpha$, l'hypothèse nulle est rejetée et inversement.

son des deux méthodes précédentes. En premier temps, le modèle fait intervenir un certain nombre de variables, puis la méthode procède à l'élimination et l'ajout successif des variables en fonction de leurs contribution dans l'amélioration du modèle.

Dans cette étude, la méthode de sélection utilisée est la *Stepwise*.

L'amélioration de la qualité du modèle est mesurée à l'aide de critère d'information d'*Akaike AIC*. Le modèle final sélectionné comprend environ 20 variables explicatives jugées comme pertinentes.

### Application du modèle

La méthode de sélection *StepWise* a permis de limiter le nombre de variable introduites dans le modèle. Le modèle final est appliqué sur chacune des trois bases. Chaque base contient la liste des variables retenues par la *StepWise*. La méthode de validation croisée utilisée est *k-fold cross-validation* avec k égale à 5.

La modèle est calibré sur la base d'apprentissage en utilisant la technique de ré-échantillonnage *DownSampling*. Une comparaison des résultats de deux méthodes de ré-échantillonnage est présentée dans la section suivante.

Rappelons que la mesure de performance des modèles et l'aire sous la courbe *ROC*. Notre objectif est de maximiser la sensibilité et la spécificité du modèle, autrement dit, nous cherchons à diminuer le nombre de faux positifs (les fausses alertes) et le nombre de faux négatifs.

Le tableau suivant récapitule les résultats de la calibration de la régression logistique sur les trois bases d'apprentissage. *L'Accuracy* correspond au taux de bon classement du modèle.

| Base | Accuracy | Sensibilité | Spécificité | AUC |
|---|---|---|---|---|
| Vol | 72% | 67% | 77% | 78% |
| Incendie | 60% | 58% | 62% | 66% |
| Circulation | 66% | 62% | 70% | 71% |

TABLE 4.1 – Bilan des résultats de la régression logistique sur la base d'apprentissage

### Comparaison de méthodes d'échantillonnage

Pour améliorer le pouvoir prédictif des modèles et vu le nombre faible de cas de fraude dans la base, nous avons testé différentes techniques de

ré-échantillonnage, notamment le sous-échantillonnage *DownSampling* qui consiste à diminuer le nombre des cas non frauduleux et la méthode *Smote* qui correspond à la génération artificielle des individus synthétiques ce qui conduit à l'augmentation du nombre de cas de fraude.

Les deux méthodes de ré-échantillonnage sont appliquées aux différentes bases d'études. Or, la méthode de sous-échantillonnage peut induire à un risque de sur-apprentissage. Par conséquent, la comparaison des deux méthodes est faite sur la base de validation.

Les résultats obtenus sont affichés dans le tableau ci-dessous.

| Base | AUC | |
|---|---|---|
| | DownSampling | SMOTE |
| Vol | 82% | 76% |
| incendie | 65% | 64% |
| Circulation | 70% | 69% |

TABLE 4.2 – Comparaison des méthodes de ré-échantillonnage

La méthode *DownSampling* apporte globalement une légère amélioration du pouvoir prédictif du modèle par rapport à la méthode *Smote*.

La méthode retenue pour la calibration des autres modèles et la présentation des résultats finaux est le *DownSampling*. En effet, cette solution nous semble la plus pertinente car elle permet également de diminuer le temps d'exécution des modèles et essentiellement la recherche des hyper-paramètres pour les modèles paramétrés.

### Choix du seuil

Pour attribuer une classe à une probabilité, le seuil *cutoff* est fixé par défaut à 0.5 pour la majorité des modèles. Si la probabilité est supérieure à 0.5, alors l'individu sera classé comme positif. Le seuil est optimisé en ayant un compromis entre le taux de faux positifs et le taux de faux négatifs. Si le taux de faux positifs est très élevé, le seuil doit être augmenté pour la classe positive et diminué pour la classe négative.

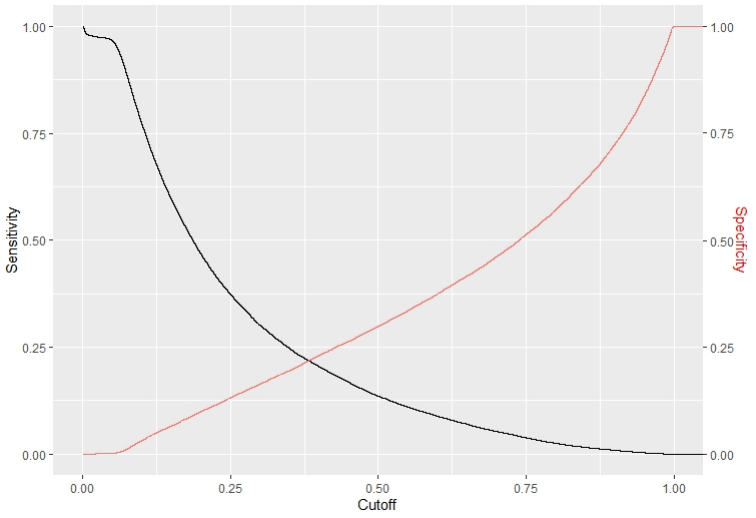



FIGURE 4.1 – Choix du seuil de probabilité

Cette figure montre la façon optimale de choisir le seuil *cutoff* pour la base *vol*. Ce seuil est fixé alors à 0.35.

### Importance des variables

Pour mesurer la contribution de la variable dans l'explication de la variable réponse, l'analyse de la courbe ROC est conduite. Pour les variables catégorielles, l'évaluation de l'importance de la variable revient à mesurer l'impact de chacune de ces modalités. Nous avons normalisé la mesure à 100 par rapport à la première variable et les variables sont ordonnées par ordre d'influence décroissante.

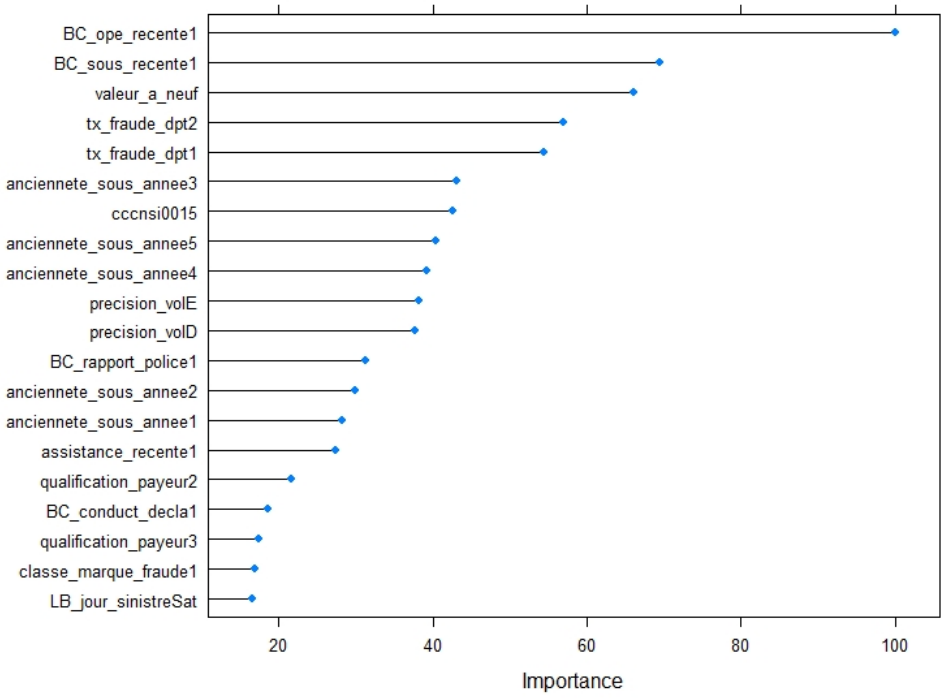



FIGURE 4.2 – Importance des variables par régression logistique sur la base *vol*

La variable *BC_ope_recente1* apparait en tête de la liste. Elle correspond à la modalité (1) autrement « oui » de la variable *BC_ope_recente*. Cette variable permet de savoir si une opération (avenant) récente a été effectuée avant le sinistre, principalement l'ajout d'une garantie.

La variable *BC_sous_recente*1 correspond à la modalité (1) autrement « oui » de la variable *BC_sous_recente*. Elle indique si la souscription est récente, c'est-à-dire, le délai entre la date de souscription et la date de survenance de sinistre est inférieur à six mois.

Nous rappelons que les experts accordent une grande importance à ces indicateurs. En règle générale, le risque de fraude semble plus important en présence de ces indicateurs.

La variable *precision_vol* est associée à la garantie vol. Elle permet de

préciser si le véhicule est non retrouvé, retrouvé avant indemnisation ou bien retrouvé après indemnisation.

La variable *cccnsi* définit les circonstances du sinistre. En effet, la base *vol* contient les sinistres dont l'évènement associé peut être le vol, une tentative de vol ou bien le vandalisme.

La variable *valeur_a_neuf* est très importante et ressort pour les trois bases. En effet, cette corrélation entre le remboursement en valeur à neuf et les tentatives de fraude est constatée également par les experts.

Nous avons choisi de présenter les résultats de la modélisation sur la base *vol*. Cette procédure a été menée pour toutes les bases et nous avons noté peu de différence au niveau des variables explicatives sélectionnées. (Voir annexe 6)

### Validation du modèle

Ce tableau récapitule les résultats de l'application du modèle sur la base de validation. Cette dernière constitue 20% des observations de chacune des bases initiales (*circulation*, *vol* et *incendie*). D'autres mesures ont été calculées afin d'évaluer les performances du modèle.

| Base | Accuracy | Sensibilité | Spécificité | F1 mesure | AUC |
|---|---|---|---|---|---|
| Vol | 76% | 71% | 77% | 20% | 82% |
| Incendie | 63% | 57% | 66% | 48% | 65% |
| Circulation | 71% | 62% | 71% | 10% | 71% |

TABLE 4.3 – Bilan final des résultats de la régression logistique

La régression logistique a permis d'avoir une première vision sur les variables qui influencent la probabilité qu'un sinistre soit frauduleux.

Globalement, les résultats obtenus sont satisfaisants. Pour la base *incendie*, la performance du modèle est moins importante. Ceci s'explique par le faible nombre de sinistres dans cette base et par conséquent le faible nombre de cas de fraude.

## 4.3 Les forêts aléatoires

Contrairement à la régression logistique, l'algorithme proposé par Breiman est une méthode paramétrique. Afin d'être calibré, les paramètres du modèle *Random Forest* doivent être fixés. Ce modèle présente alors trois paramètres principaux :

- Le nombre de variables explicatives choisies aléatoirement dans chaque arbre noté q. Rappelons que, par défaut, ce nombre est fixé à $\sqrt{p}$ où p est le nombre de variables explicatives dans la base.
- Le nombre d'arbres à construire dans la forêt. Sa valeur par défaut est fixée à 500.
- Le nombre minimum d'observations dans chaque nœud terminal permettant de contrôler la complexité de l'arbre.

Comme il s'agit d'un modèle paramétré, il est judicieux de choisir les paramètres optimaux pour obtenir un modèle performant. Le but de l'optimisation des paramètres est d'éviter les phénomènes de sur-apprentissage et de sous-apprentissage. Les hyper-paramètres sont les paramètres à définir pour ajuster un modèle afin d'avoir le meilleur score. La procédure de la recherche des hyper-paramètres s'appelle le *Tuning*. Cette approche consiste à construire une grille ou un quadrillage (*Grid search)* contenant une liste des valeurs possibles de chaque paramètre dans le modèle. Le modèle est entrainé pour chacune des combinaisons et son score est enregistré. Par exemple, ce score correspond à l'erreur *Out-of-Bag*. Le modèle ayant la valeur de l'erreur *Out-of-Bag* la plus faible est choisi et les paramètres de ce modèle correspondent au meilleur paramétrage. Bien que cette technique soit performante, elle est couteuse en terme de temps.

### Calibration du modèle

Les trois bases comportent environ une trentaine de variable. Par défaut, la valeur de *mtry* est fixée à 5 et le nombre d'arbres est fixé à 500 arbres [3].

Le modèle est calibré sur les trois bases d'apprentissage en construisant une grille de paramètres. La recherche des hyper-paramètres s'accompagne aussi de la procédure de validation croisée, notamment la méthode *k-fold cross-validation* avec k fixé à 5.

La grille conçue comporte une liste de valeurs pour chaque paramètre. Le nombre d'arbres à agréger varie entre 100 et 1000, pour chaque nombre

3. Pour construire le modèle, nous utilisons la librairie Caret fournie par le logiciel R.

d'arbres, une valeur de *mtry* est testée. La sélection se base sur la minimisation de l'erreur *Out-of-bag*.

Ce graphique donne un aperçu sur le *tuning* des hyper-paramètres. Nous avons choisi de présenter l'évolution de *l'AUC* en fonction des paramètres du modèles.

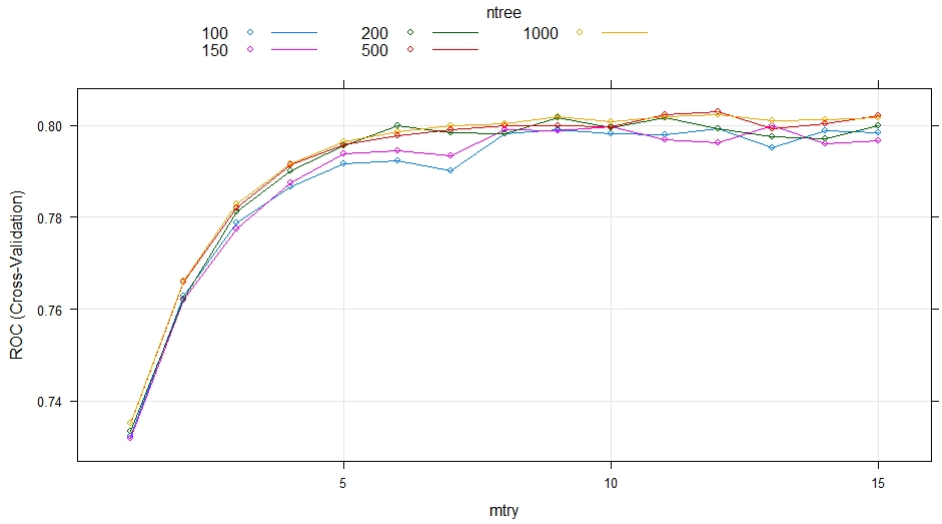



FIGURE 4.3 – Tuning des forêts aléatoires

Les résultats affichés concernent la base *vol*. La recherche des hyper-paramètres a permis d'obtenir un *mtry* optimal égal à 12 et un nombre d'arbres égal à 500. L'erreur *OOB* est estimée à 26%.

L'exercice est répété pour chacune des trois bases d'apprentissage, le tableau ci-dessous montre les résultats obtenus.

| Base | Accuracy | Sensibilité | Spécificité | AUC |
|---|---|---|---|---|
| Vol | 74% | 68% | 80% | 83% |
| Incendie | 62% | 60% | 65% | 68% |
| Circulation | 67% | 62% | 73% | 73% |

TABLE 4.4 – Calibration des Forêts aléatoires

### Validation du modèle

Les prédictions ont été faites sur la base de validation qui constitue 20% de la base initiale. Les paramètres des trois modèles correspondent aux paramètres optimaux trouvés par le tuning et identiquement à la régression logistique, un seuil (*cutoff*) est choisie pour classer les probabilités . Ce tableau résume les résultats obtenus.

| Base | Accuracy | Sensibilité | Spécificité | F_mesure | AUC |
|---|---|---|---|---|---|
| Vol | 78% | 73% | 79% | 22% | 83% |
| Incendie | 64% | 59% | 65% | 49% | 67% |
| Circulation | 73% | 64% | 73% | 9% | 74% |

TABLE 4.5 – Bilan des résultats des forêts aléatoires

Pour la base *circulation*, le nombre de cas de fraude est très faible par rapport au nombre total des observations. La technique de ré-échantillonnage a permis d'améliorer la performance de modèle. En effet, l'application du modèle n'était pas possible sur la base *circulation* initiale en raison du nombre important des sinistres et du faible nombre de cas de fraude.

### Importance des variables

La sélection des variables importantes se base sur le critère *mean decrease gini*. La liste des variables les plus discriminantes pour la base *vol* est présentée dans le graphique ci-dessous.

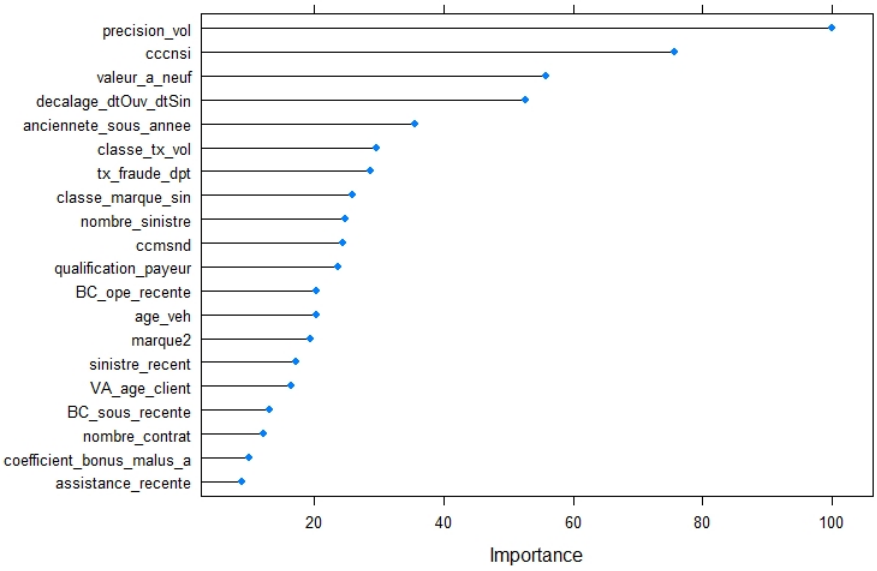



FIGURE 4.4 – Importance des variables par forêts aléatoires sur la base *vol*

Les variables importantes qui ressortent sont quasiment similaires à celles trouvées par la régression logistique. C'est la variable *précision vol* qui arrive en tête de la liste. Cette variable indique si le véhicule a été retrouvé suite à la déclaration du vol. Elle est suivie de la variable *cccnsi* qui renseigne sur les circonstances du sinistre. La *valeur à neuf* apparait dans le top 3 semblablement au modèle précédent.

La variable *decalage_dtOuv_dtSin* correspond au décalage entre la date d'ouverture et la date de survenance du sinistre. Plus ce délai est élevé, plus le risque de frauder est important.

La variable *classe_tx_vol* informe sur la classe de taux de vol des véhicules par département. Ce taux est publié par le ministère d'intérieur. Trois modalités peuvent être distinguées : un taux faible (classe 0), un taux moyen (classe 1) et un taux élevé (classe 2).

Nous retrouvons aussi d'autres variables comme *BC_ope_récente*, *sinistre_récent*, *BC_sous_récente* et *assistance_recente* qui correspondent respectivement aux

indicateurs métiers opération récente, sinistralité récente, souscription récente et assistance récente.

La qualification payeur qui renseigne sur la qualité du paiement de l'assuré contribue à l'explication de la fraude.

La variable *ccmsnd*[4] donne une information sur le motif du déplacement au moment du sinistre.

## 4.4 XGBoost

L'algorithme *XGBoost* propose plusieurs paramètres. Certains sont essentiels et montrent une forte influence sur les performances du modèle. Parmi ces paramètres :

- **nrounds** est le nombre d'itérations à effectuer. Plus la valeur est grande, plus le modèle est lent.
- **max_depth** correspond à la profondeur de l'arbre maximal. Une grande valeur de *max_depth* conduit à un modèle trop complexe et engendre un phénomène de sur-apprentissage. En contre partie, une valeur faible de *max_depth* augmente le risque du sous-apprentissage. La valeur par défaut est fixée à 6.
- **colsample_bytree** est le pourcentage des variables choisies aléatoirement parmi l'ensemble de tous les attributs au moment de la construction de l'arbre. C'est l'équivalent de *mtry* dans l'algorithme de Random Forest.
- **eta** (*learning rate*) est un paramètre qui contrôle les poids des arbres conçus dans le modèle. Par défaut il est fixé à 0.3.
- **subsample** détermine le pourcentage des observations à utiliser pour construire l'arbre. Par défaut, ce paramètre est égal à 1.
- **gamma** détermine la réduction minimale des pertes (la fonction cout) requise pour effectuer une partition supplémentaire sur un nœud terminal de l'arbre. Une grande valeur conduit à un modèle plus conservateur / prudent. Sa valeur par défaut est fixée à 0.

4. Voir Annexe 6 pour la liste des variables

### Calibration du modèle

La calibration du modèle *XGBoost* est similaire à celle des forêts aléatoires. Une grille de hyper-paramètres est conçue. Nous cherchons alors à maximiser *l'AUC*. La sensibilité et la spécificité du modèle sont contrôlées afin de diminuer le nombre de fausses alertes générées.

| Base | Accuracy | Sensibilité | Spécificité | AUC |
|---|---|---|---|---|
| Vol | 74% | 68% | 80% | 80% |
| Incendie | 61% | 61% | 61% | 66% |
| Circulation | 67% | 61% | 74% | 73% |

TABLE 4.6 – Calibration du modèle XGBoost

Vu le grand nombre de paramètres à optimiser, le tuning du modèle *XGBoost* est très couteux en terme du temps. Le choix des paramètres est fait sur plusieurs étapes par validation croisée. Dans un premier temps, nous faisons varier le nombre d'itérations *nround* entre 100 et 1000 [5] et la profondeur de l'arbre entre 5 et 20 pour deux valeurs de gamma 0 et 1. Une fois, le nombre d'itérations optimal est choisi, nous faisons varier les autres paramètres essentiellement le *shrinkage eta* qui est responsable de la prévention contre le sur-apprentissage. (Voir Annexe 5)

### Validation du modèle

| Base | Accuracy | Sensibilité | Spécificité | F_mesure | AUC |
|---|---|---|---|---|---|
| Vol | 78% | 75% | 79% | 23% | 83% |
| Incendie | 64% | 62% | 64% | 50% | 66% |
| Circulation | 74% | 62% | 74% | 9% | 74% |

TABLE 4.7 – Bilan des résultats de XGBoost

Plusieurs observations appartenant à la classe des négatifs sont prédits comme des positifs. C'est pour cette raison, la *F_mesure* est faible pour la base circulation.

5. ou 500 pour la base *circulation*, vue sa taille très élevée

### Importance des variables

De la même manière que les autres modèles, une des sorties du modèle est l'importance des variables. Elle permet d'identifier les variables pertinentes dans la modélisation de la fraude.

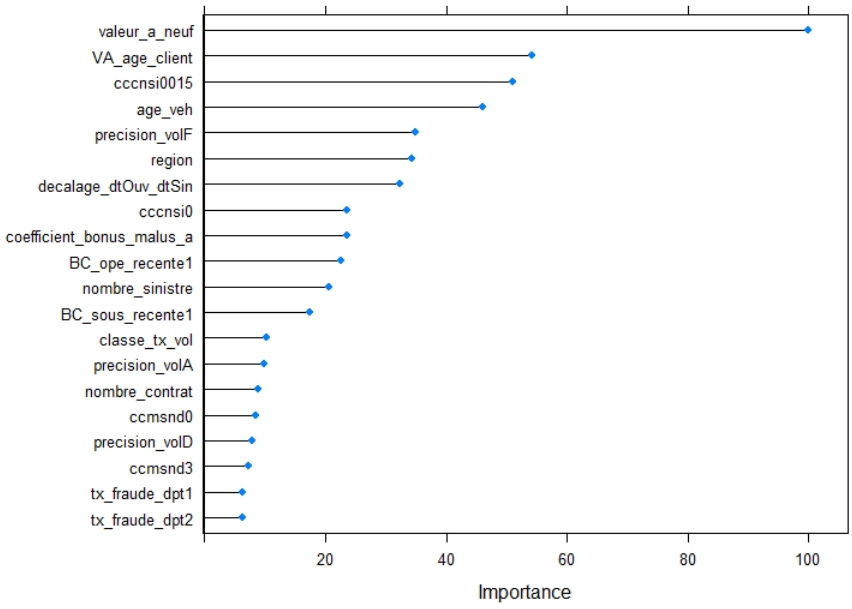



FIGURE 4.5 – Importance des variables par XGBoost pour la base *vol*

Globalement, les variables estimées importantes par le *gradient boosting* sont quasiment les mêmes que celles obtenues par les forêts aléatoires. L'ordre d'importance des variables est également semblable à celui du modèle précédent.

Chapitre 5

# Analyse des résultats et perspectives

## 5.1 Bilan et validation des modèles

Afin d'évaluer les résultats des différents modèles, nous faisons un bilan de performance de chaque modèle ainsi que ses avantages et ses inconvénients. De plus, le temps de calcul et d'exécution constitue un point important dans le choix du modèle.

### 5.1.1 Avantages et inconvénients des modèles

#### Régression Logistique

**Avantages**

— Un modèle non paramétré.

— Facilité d'interprétation

— Présence de coefficients et aussi de la direction de l'association (coefficient positif ou négatif)

**Inconvénients**

— Transformation des modalités des variables catégorielles en variables qualitatives

— Problème d'optimisation plus complexe

— Sensible aux corrélations
— Absence de détection des relations non linéaires

### Forêts aléatoires

**Avantages**

— Un modèle robuste face aux données aberrantes et manquantes
— Le preprocessing des variables n'est pas nécessaire (la normalisation des données n'est pas requise)
— Le nombre de paramètres à optimiser n'est pas élevé
— Possibilité de parallélisation
— Pas de risque de sur-apprentissage
— Échantillonnage intégré
— Traitement des variables quantitatives et qualitatives

**Inconvénients**

— Temps de calcul lent
— Nécessité de paramétrage
— L'interprétation du modèle est moins évidente.

### Extreme Gradient Boosting : XGBoost

**Avantages**

— Absence du risque de sur-apprentissage (avec le paramètre de la régularisation)
— Parallélisation
— Traitement des données manquantes
— Gestion des interactions entre les variables
— Capacité de détection des relations non linéaires entre les données

**Inconvénients**

— Complexité d'interprétation : L'algorithme est qualifié de « boite noire »
— Temps de calcul élevé
— Un nombre important des paramètres à définir

### 5.1.2 Comparaison des résultats

La première sélection de variables a été effectuée à travers un modèle *Xgboost* sans une éventuelle optimisation des paramètres. Ensuite, les trois modèles sont appliqués sur les trois bases.

le tableau suivant récapitule les résultats des modèles sur chacune des bases de validation.

| | AUC | | |
|---|---|---|---|
| Bases | Régression logistique | Forêts aléatoires | XGBoost |
| Vol | 82% | 83% | 83% |
| Incendie | 65% | 67% | 66% |
| Circulation | 71% | 74% | 74% |

TABLE 5.1 – Comparaison des résultats

Nous remarquons que la performance des trois modèles est globalement satisfaisante ($AUC$ est de l'ordre de 70%). Les résultats sont proches. Certes, nous constatons une amélioration du pouvoir prédictif lors de l'utilisation des modèles basés sur la construction des arbres, du fait que ces modèles sont capables de détecter les relations non linéaires entre les variables.

Le principal inconvénient des algorithmes basés sur la construction des arbres est la complexité à interpréter ses résultats. En effet, avec ces modèles, la performance est améliorée au détriment de *l'interprétabilité*. Néanmoins, il est nécessaire de comprendre la décision d'un modèle à classer un dossier comme frauduleux pour pouvoir communiquer le résultat aux experts métiers.

Les équipes métiers ont besoin de comprendre le résultat du modèle afin de le valider et d'acter son utilité. Nous avons créé de nombreuses variables à partir des *règles métier* afin de rapprocher ou d'affecter un dossier à un indicateur. Ces indicateurs constituent des pistes d'investigation pour les experts. Or, l'exercice n'est pas facile et demeure compliqué à implémenter.

## 5.2 Perspectives et limites

### 5.2.1 Critiques et améliorations envisageables des modèles

Dans ce mémoire, nous avons présenté trois modèles d'apprentissage supervisé pour déterminer les variables les plus discriminantes et les plus explicatives de la fraude. Les résultats obtenus par ces méthodes sont proches

et globalement satisfaisants. Néanmoins, plusieurs pistes d'amélioration sont envisageables afin d'avoir une meilleure performance.

La liste des variables sélectionnées dans les modèles n'est bien évidemment pas exhaustive et ne contient pas toutes les informations disponibles sur la déclaration des sinistres. Afin d'améliorer la modélisation de la fraude, il serait intéressant d'intégrer d'autres variables dans le modèle comme le montant du sinistre, la prime, le nombre de personnes impliquées dans le sinistre... Ces informations n'ont pas été utilisées, soit parce qu'elles ne figuraient initialement pas dans les bases mises à disposition, soit parce que leurs récupération n'était pas un exercice évident.

Dans cette étude, notre objectif était de réduire au minimum l'erreur de prédiction des modèles. Pour ce faire, nous avons choisi *l'AUC* comme critère d'évaluation. Cette mesure a été complétée par un contrôle de la matrice de confusion en évaluant la sensibilité et la spécificité.

Les suites de notre travail pourraient également se pencher sur l'expérimentation d'autres mesures d'évaluation, notamment l'*AUC Précision-Rappel AUC PR* qui est différent de *l'AUC ROC*. C'est l'aire normalisée sous la courbe paramétrique définie par la précision et le rappel en fonction du seuil de décision. La précision et le rappel se focalisent sur la classe des positifs et ne tiennent pas compte des vrais négatifs (la classe majoritaire).

Une troisième solution qui pourrait améliorer les modèles utilisés est de faire appel au *Stacking*. Il s'agit de combiner plusieurs algorithmes afin de concevoir un modèle plus robuste et plus performant.

Par ailleurs, les résultats obtenus dépendent principalement de la base *StatFraude* qui n'est évidemment pas alimentée d'une manière exhaustive. En effet, la démonstration de la fraude est le produit d'une analyse humaine basée sur la constatation d'un scénario, la fraude n'existe que si le conseiller l'a préalablement identifiée et analysée. Ceci constitue le principal biais dans cette étude et permet de mettre en évidence la nécessité d'un investissement *métier* préalable pour améliorer l'échantillon.

### 5.2.2 Perspectives

La recherche de fraude est aujourd'hui basée sur l'expertise. Par conséquent, l'échantillon des fraudes avérées n'est pas suffisamment important pour pouvoir construire des modèles fiables et robustes.

Dans cette étude, nous avons opté à la modélisation de la fraude à l'aide des algorithmes d'apprentissage supervisé. Cette catégorie nécessite la présence des données labellisées afin de concevoir un modèle de prédiction. Or, il existe une deuxième catégorie d'apprentissage automatique qui est

l'apprentissage non supervisé. Il s'agit de détecter les similarités dans les données non étiquetées et de créer ainsi des classes ou bien des groupes d'individus homogènes présentant des caractéristiques similaires et communes.

Par ailleurs, la détection d'anomalies constitue une catégorie d'apprentissage non supervisé. Cette méthode permet de résoudre des problèmes liés à la fraude. En effet, en raison du faible nombre de cas de fraudes avérées, la fraude est considérée comme une anomalie ou un évènement rare.

Plusieurs articles font le point autour de la modélisation de la fraude en apprentissage non supervisé. Les deux modèles identifiés comme les plus adaptés à la détection de la fraude sont *Isolation Forest*[1] et *DBScan*[2].

Dans cette optique, nous avons expérimenté ces deux algorithmes. Pour évaluer leurs performances, nous avons utilisé le référentiel *StatFraude* afin de comparer les sinistres identifiés comme anomalie aux fraudes avérées. Cependant, les résultats n'étaient pas satisfaisants. Plusieurs hypothèses se manifestent pour expliquer ce résultat. La principale raison demeure la non exhaustivité du référentiel *StatFraude*. La base *StatFraude* est le produit d'une activité des experts et la constitution de ce fichier se base principalement sur les *règles métier*. Il se pourrait que de nombreux cas de fraudes aient été non détectés et les données ne soient pas systématiquement renseignées.

De ce fait, nous nous sommes focalisés exclusivement sur les algorithmes d'apprentissage supervisé. Ces derniers peuvent être complétés par une approche non supervisée ou semi-supervisée. Cette combinaison pourrait être efficace pour améliorer la détection des fraudes.

1. C'est un algorithme de partitionnement basé sur les arbres de décision. Il se base sur le calcul d'un score d'anomalie pour chaque observation en l'isolant d'une récursive des autres données.

2. Density-Based Clustering of Applications with Noise. Cet algorithme se base sur la densité ds points pour établir des clusters et identifier les valeurs aberrantes.

# Conclusion

L'objectif de ce travail était de développer une approche de détection des fraudes potentielles en assurance automobile basée sur des algorithmes mathématiques et des méthodes d'apprentissage statistique. La première partie de cette étude a consisté à modéliser la fraude en utilisant des algorithmes d'apprentissage supervisé.

La régression logistique, les forêts aléatoires, les algorithmes de gradient *boosting* constituent les méthodes d'apprentissage supervisé introduites dans ce mémoire. Ces méthodes ont été calibrées sur trois bases de sinistres automobile de la marque $MAAF$ du groupe *Covéa*, construites en fonction de l'évènement du sinistre.

Le principal problème rencontré dans la modélisation de la fraude par *Machine Learning* est la faible proportion des sinistres frauduleux par rapport au nombre total de sinistres. Afin d'équilibrer les données et d'éviter le risque de sur-apprentissage, nous avons effectué un ré-échantillonnage de la base et nous avons opté pour la validation croisée. Cette approche a été précédée par une analyse descriptive de la base et une étude de corrélation entre les variables. Pour évaluer le pouvoir prédictif des modèles, nous avons utilisé l'aire sous la courbe $ROC\ AUC$.

Pour les trois modèles, les résultats étaient proches avec un $AUC$ de plus de 65%. De plus, les variables jugées importantes dans la modélisation de la fraude étaient quasiment les mêmes pour les trois modèles et ont montré une cohérence avec les *règles métier*.

L'application des méthodes d'apprentissage supervisé nécessite l'existence des données labellisées, c'est-à-dire, une base dotée d'une variable per-

mettant de classifier les sinistres en tant que sinistres frauduleux ou non. Or, la base dont dispose *Covéa* n'est pas un référentiel exhaustif de tous les cas de fraude, mais une consolidation de l'activité des équipes opérationnelles. Pour pallier à ce problème de fiabilité de la variable indiquant la fraude, plusieurs solutions sont envisageables comme les méthodes d'apprentissage non supervisé et semi-supervisé.

Pour conclure, cette étude constitue les premiers travaux de modélisation de la fraude à travers une approche *data science* au sein de *Covéa*. Ce projet a démontré la capacité du groupe à détecter la fraude sur la base des algorithmes d'apprentissage statistique. De ce fait, *Covéa* souhaite renforcer ses moyens de détection de fraude en améliorant son système d'information, en l'occurrence la qualité des données nécessaires dans la modélisation, en intégrant les démarches de recherche de fraude dès la souscription du contrat, et également en se dotant des technologies appropriées.

La détection de fraude est devenue un enjeu stratégique pour l'assureur dont objectif est de disposer d'un système efficace et automatisé en matière de lutte contre la fraude.

# Bibliographie


[1] BAPTISTE GREGORUTTI. Forêts aléatoires et sélection de variables. 2015.

[2] ROBIN GENUER & JEAN-MICHEL POGGI. Arbres CART et Forêts aléatoires, Importance et sélection de variables. 2016. hal-01387654v1

[3] RICCO RAKOTOMALALA. Pratique de la Régression Logistique, Régression Logistique Binaire et Polytomique. Version 2.0. 2015

[4] SAMEH BORGI. Une analyse économique et expérimentale de la fraude à l'assurance et de l'audit. 2006

[5] LAURENT ROUVIÈRE. Introduction aux méthodes d'agrégation : boosting, bagging et forêts aléatoires.

[6] Rapport moral ALFA. 2018. *Document interne confidentiel.*

[7] JONATHON KARSENTY. La détection de fraudes à l'assurance. 2016

[8] REKHA BHOWMIK. Detecting Auto Insurance Fraud by Data Mining Techniques. University of Texas at Dallas, USA. 2011

[9] CAROL HARGREAVES & VIDYUT SINGHANIA. Analytics for Insurance Fraud Detection : An Empirical Study. National University of Singapore. 2016

[10] IGOR ANOHHIN. Data Mining And Machine Learning for Fraud Detection. Master's thesis. Tallinn University of Technology. 2017

[11] RÉMI DOMINGUES. Machine Learning for Unsupervised Fraud Detection. Royal Institute of Technology. 2015

[12] KATJA MULLER. The Identification of Insurance Fraud : An Empirical Analysis. University of St. Gallen. Working Papers on Risk Management and Insurance No.137. 2013

[13] DELOITTE CONSULTING LLP. Predictive Modeling for Life Insurance. 2010

[14] ATLAS MAGAZINE. Fraude à l'assurance : le coût et la lutte contre ce fléau. 2017. https ://www.atlas-mag.net/article/fraude-a-l-assurance-le-cout-et-la-lutte-contre-ce-fleau

[15] AQUILA DATA ENABLER. Détection de fraude, comportements atypiques et détection d'anomalie. 2018

[16] ARGUS DE L'ASSURANCE. Lutte contre la fraude : de l'intention à la réalité. 2018

[17] ERCOM. Fraude et interprétabilité des modèles de Machine Learning.

[18] RIYA ROY & K. THOMAS GEORGE. Detecting insurance claims fraud using machine learning techniques. (ICVES) IEEE International Conference, 2017

[19] WIKISTAT. Agrégation des modèles. Institut mathématique de Toulouze

[20] WIKISTAT. Régression logistique ou modèle binomial. Institut mathématique de Toulouze

[21] L.BREIMAN. Bagging predictors. Machine Learning 26. 1996

[22] L.BREIMAN. Random Forests. Machine Learning 45. 2001

[23] Y.FREUND & R.E.SCHAPIRE. Experiments with a new boosting algorithm. Machine Learning : proceedings of the Thirteenth International Conference. 1996

[24] T.HASTIE, R.TIBSHIRANI & J FRIEDMAN. The elements of statistical learning : data mining, inference and prediction, Springer. 2009

[25] R.SCHAPIRE. The boosting approach to machine learning. An overview MSRI workshop on non linear estimation and classification. 2002

[26] R.GUHA, SHREYA MANJUNATH & KARTHEEK PALEPU. Comparative analysis of Machine Learning techniques for detecting insurance claims fraud.

[27] RICCO RAKOTOMALA. Analyse de corrélation. Étude des dépendances. 2017

[28] G.LOUPPE, L.WEHENKEL, A.SUTERA & P.GEURTS. Understanding variable importances in forests of randomized trees. 2013

[29] FRÉDÉRIC NGUYEN KIM. La détection de la fraude en vol automobile, un enjeu important. Argus de l'assurance. 2011

[30] CHARLES LE CORROLLER. Le contrat d'assurance. 2017

[31] PHILIPPE BESSE. Apprentissage statistique. 2016

[32] ANTOINE LY. Algorithmes de machine learning en assurance : solvabilité, textmining, anonymisation et transparence. 2019

[33] GODFRIED GUENDEHOU. Construction d'un score temporel pour la détection des sinistres frauduleux en assurance automobile par apprentissage statistique. Synthèse de mémoire. 2017

# Annexes

## Annexe 1 : Régression logistique

Dans cet annexe, nous montrons que les deux approches présentées dans la partie régression logistique sont équivalentes. En effet :

$$ln[\frac{\pi}{1-\pi}] = a_0 + a_1X_1 + ... + a_nX_n$$

$$\ln[\frac{P(X|Y=1)}{P(X|Y=0)}] = b_0 + b_1X_1 + ... + b_nX_n$$

avec

Y est la variable cible, X est le vecteur des variables explicatives et $P(Y=1|X) = \mathbb{E}(Y|X) = \pi$

$$\begin{aligned} ln[\frac{\pi}{1-\pi}] &= a_0 + a_1X_1 + ... + a_nX_n \\ &= ln[\frac{P(Y=1)}{P(Y=0)} * \frac{P(X|Y=1)}{P(X|Y=0)}] \\ &= ln[\frac{P(Y=1)}{P(Y=0)}] + ln[\frac{P(X|Y=1)}{P(X|Y=0)}] \\ &= ln[\frac{P(Y=1)}{P(Y=0)}] + (b_0 + b_1X_1 + ... + b_nX_n) \end{aligned}$$

et donc $a_0 = ln[\frac{P(Y=1)}{P(Y=0)}] + b_0$

### Test de Wald

Le test de Wald teste l'hypothèse nulle selon laquelle un paramètre est égal à la valeur zéro.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

La statistique du test de Wald :

$$\frac{\hat{\beta}_j}{s_{\beta_j}} \sim N(0,1)$$

alors

$$\frac{\hat{\beta}_j}{s^2_{\beta_1}} \sim \chi^2(1)$$

Décision de rejeter $H_0$ au niveau de risque $\alpha$ : On rejete $H_0$ si :
$Wald \geq \chi^2_{1-\alpha}(1)$ et $p - value \leq \alpha$,i.e. Prob$(\chi^2_{1-\alpha}(1) \geq Wald) \leq \alpha$

## Annexe 2 : Arbre de décision CART

CART désigne Classification And Regression Trees. C'est une méthode introduite par Breiman en 1984 qui est utilisée en régression et en classification.

Le principe des arbres CART est de partitionner d'une manière récursive et binaire l'ensemble des observations en sous-groupes homogènes. L'algorithme CART s'applique en deux étapes : la première consiste à construire un arbre maximal et la deuxième concerne la phase d'élagage.

Lors de la phase de construction, l'espace des observations est partitionné en deux sous-espaces homogènes suivant une variable explicative considérée comme la variable la plus discriminante vis-à-vis la variable cible. La racine de l'arbre correspond à toutes les observations d'entrée. Les deux sous-parties homogènes constituent alors les nœuds. L'algorithme sélectionne la variable explicative qui permet d'effectuer la meilleure découpe (split). Cette variable peut être qualitative ou bien quantitative. Soit $X_j$, $j \in \{1, ..., p\}$ une variable explicative quantitative suivant laquelle la découpe a été faite (p le nombre total des variables). L'algorithme va chercher le seuil d permettant d'effectuer la meilleure segmentation. Les deux sous parties construites sont $\{X_j \leq d\}$ et $\{X_j > d\}$. La meilleure découpe est sélectionnée en minimisant la fonction de cout. En classification, il s'agit de mesurer l'impureté d'un nœud à l'aide de l'indice de Gini.

### Indice de Gini

L'indice de Gini permet de mesurer l'hétérogénéité d'un nœud N et est défini par

$$I_G(N) = \sum_i f_i(1 - f_i) = \sum_i (f_i - f_i^2) = \sum_i f_i - \sum_i f_i^2 = 1 - \sum_i f_i^2$$

Dans le cas d'un problème de classification binaire, i correspond au nombre de classe c'est-à-dire $i \in \{1, 2\}$ et $f_i$ correspond à la proportion des observations de classe i dans le nœud N.

A chaque pas de partitionnement, on cherche à augmenter l'homogénéité (diminuer l'impureté) des classes et donc à diminuer l'indice de Gini afin d'obtenir des nœuds ne contenant que des observations de même classe. Ainsi, l'arbre développé est appelé un arbre maximal et les nœuds terminaux sont appelés des feuilles. Une valeur est associée à chaque feuille et qui correspond au label de la classe majoritaire des observations définissant ce nœud.

### Élagage

La deuxième phase de l'algorithme CART consiste à améliorer les performances de l'arbre maximal voire l'optimiser. Cette phase s'appelle l'élagage. Il s'agit de construire un sous-arbre permettant de trouver un compromis biais-variance. En effet, l'arbre maximal construit possède un biais faible mais une variance importante. L'élagage correspond à la construction des sous-arbres, à partir de l'arbre maximal, (élagués de l'arbre maximal) en minimisant un critère pénalisé. L'arbre optimal est ainsi obtenu par validation croisée.

## Annexe 3 : Bagging

Soit Y une variable à expliquer, $x_1, ..., x_p$ les variables explicatives et $f(x)$ un modèle fonction.

Soit B échantillons, la prévision par agrégation de modèles est définie, en fonction de la variable à expliquer Y :

- Si Y est quantitative : $\hat{f}_B(.) = \frac{1}{B} \sum_{k=1}^{B} \hat{f}_{d^k}(.)$. Il s'agit d'une moyenne des résultats obtenus pour les modèles associés à chaque échantillon.
- Si Y est qualitative : $\hat{f}_B(.) = argmax_j \sum_{k=1}^{B} \mathbf{1}_{\hat{f_{d^k}}}(.) = j\}$. Il s'agit de voter et élire la réponse probable. Le vote majoritaire consiste à chercher la classe majoritaire parmi les classes prédites.

Le principe est élémentaire, moyenner les prévisions de plusieurs modèles indépendants permet de réduire la variance et donc de réduire l'erreur de prédiction.

Les B échantillons sont des réplications d'échantillons boostrap obtenus chacun par n tirages avec remise.

---

**Algorithme 5.1** Bagging

---

x l'observation à prévoir
$d_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ un échantillon
B le nombre d'estimateurs à agréger
Pour k =1,..., B faire :

1. Tirer un échantillon boostrap $d_n^k$ dans $d_n$.
2. Estimer $\hat{f}_{d_n^k}(x)$ sur l'échantillon boostrap.

Calculer l'estimation $\hat{f}_B(x) = \frac{1}{B} \sum_{k=1}^{B} \hat{f}_{d_n^k}(x)$ ou le résultat du vote.

---

# Annexe 4 : Corrélation

## Test de Pearson

Ce test porte sur le coefficient de corrélation linéaire $\rho$ de Bravais-Pearson pour n observations.

Le coefficient de corrélation entre deux variables aléatoires X et Y ayant chacune une variance, noté Cor(X,Y) ou $\rho_{XY}$ est défini par :

$r = \rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

où

$Cov(X, Y)$ désigne la covariance des variables X et Y

$\sigma_X$ et $\sigma_Y$ désignent leurs écarts types.

L'hypothèse nulle $H_0$ est défini par un coefficient de corrélation de Pearson nul ou les variables ne sont pas corrélées ($\rho = 0$).

On test $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$

— Si p-valeur $> 0.05$, on accepte $H_0$.

— Si p-valeur $< 0.05$, on rejette $H_0$ avec un risque de première espèce p.

## Test de Goodman et Kruskal

Ce test permet de mesurer l'association entre deux variables qualitatives, X et Y. Soit le tableau de contingence de X et Y, $n_{l,c}$ l'effectif associé à la ligne $l \in [1, L]$ et à la colonne $c \in [1, C]$, c'est l'effectif issu du croisement entre la modalité l de X et la modalité c de Y.

Le coefficient $\tau$ s'écrit :

$$\tau = 1 - \frac{n^2 - n \sum_{l=1}^{L} \sum_{c=1}^{C} \frac{n_{l,c}^2}{\sum_{c=1}^{C} n_{l,c}}}{n^2 - \sum_{c=1}^{C} (\sum_{l=1}^{L} n_{l,c})^2}$$

Si $\tau$ converge vers 0 alors la liaison entre les deux variables est absente. Si $\tau$ converge vers 1 alors la liaison entre les deux variables est forte.

## Étude des corrélations : Complément

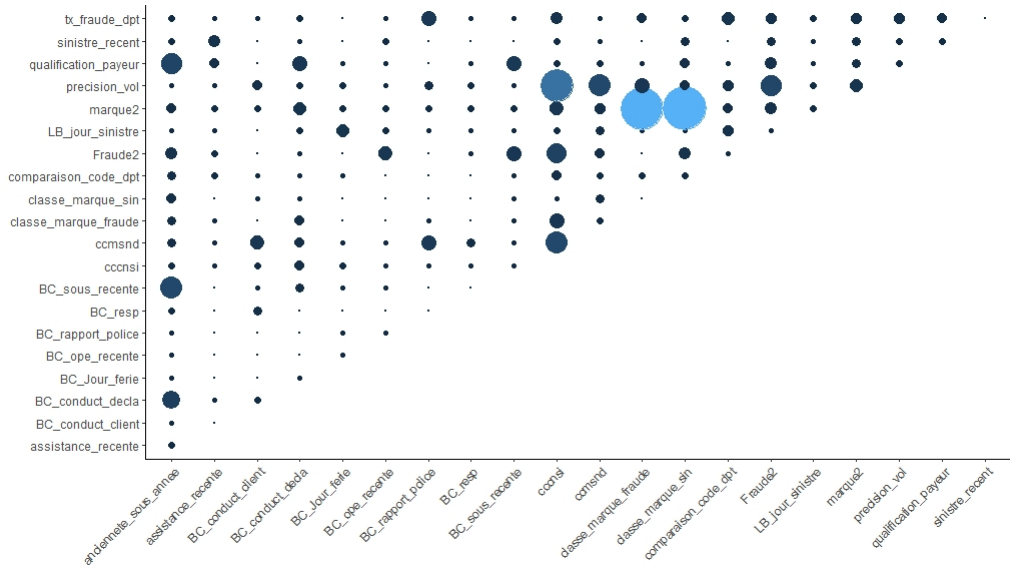



FIGURE 5.1 – Étude des corrélations dans la base *vol*

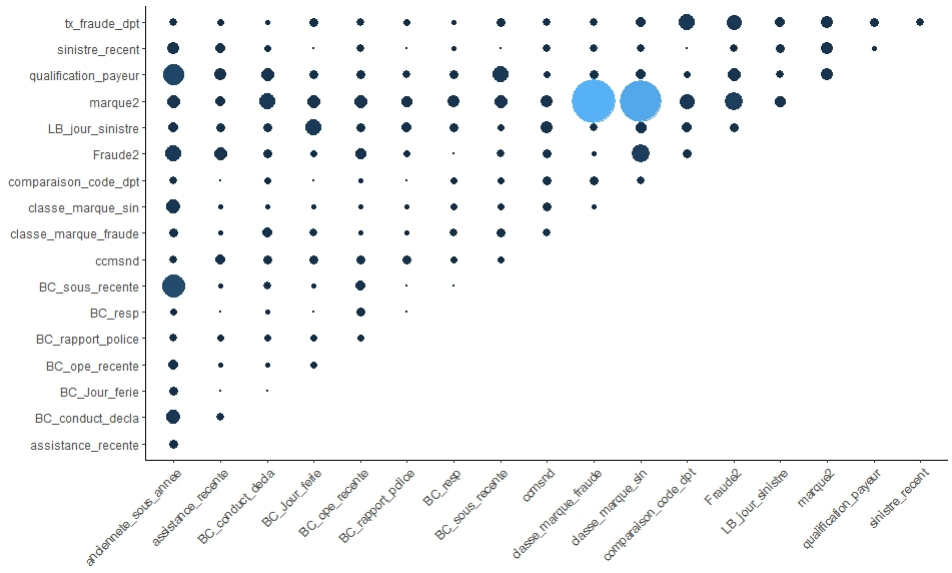



FIGURE 5.2 – Étude des corrélations dans la base *incendie*

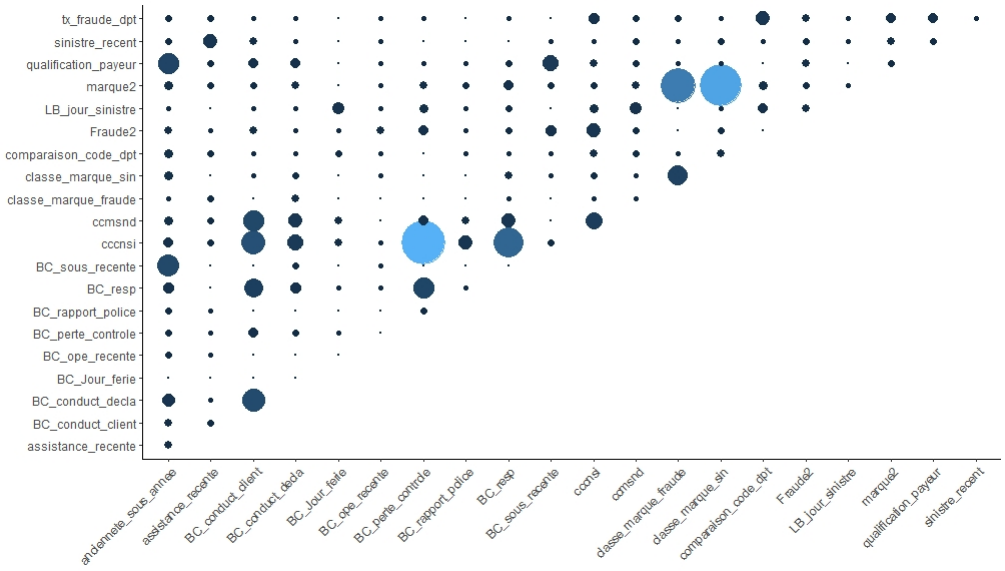



FIGURE 5.3 – Étude des corrélations dans la base *circulation*

# Annexe 5 : Résultats des modèles : Complément

## Régression logistique

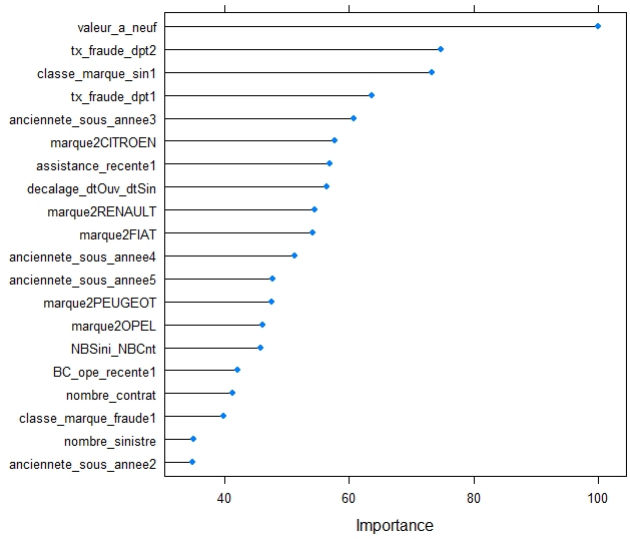



FIGURE 5.4 – Importance des variables par régression logistique pour la base *incendie*

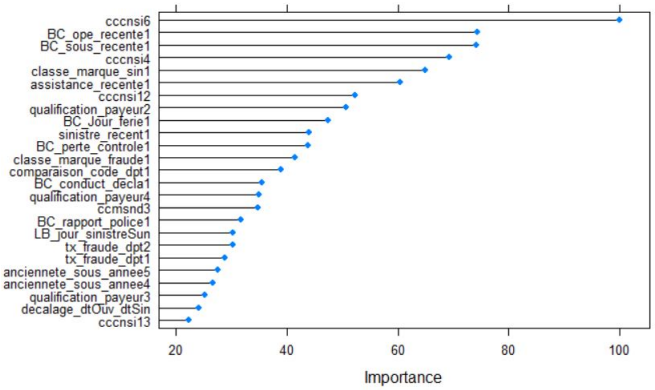



FIGURE 5.5 – Importance des variables par régression logistique pour la base *circulation*

## Random Forest

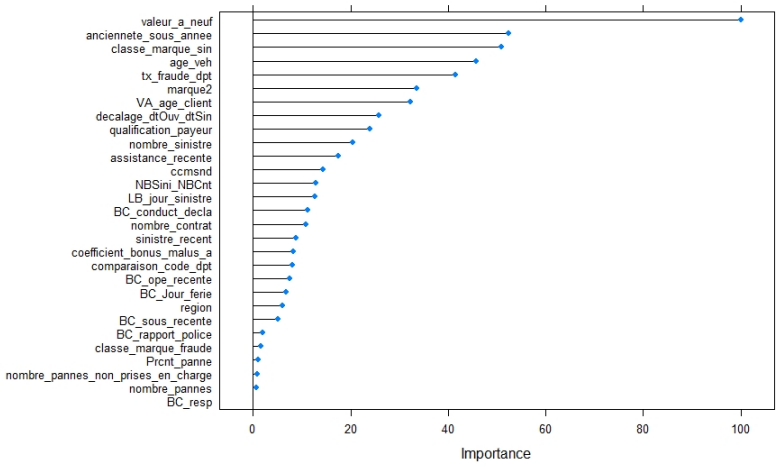



FIGURE 5.6 – Importance des variables par forêts aléatoires pour la base *incendie*

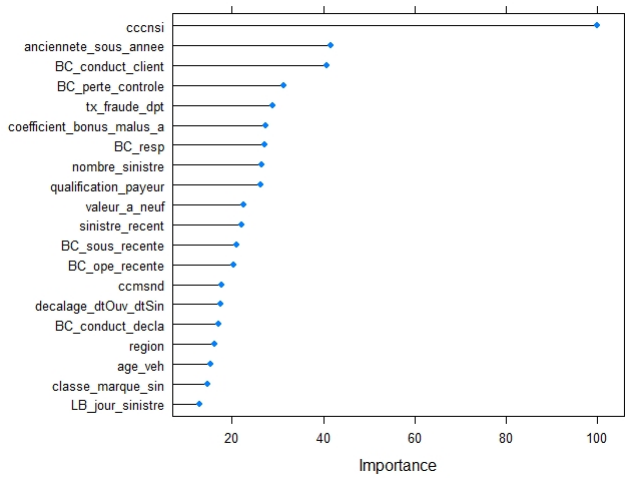



FIGURE 5.7 – Importance des variables par forêts aléatoires pour base *circulation*

## XGBoost

### Complément Tuning des paramètres

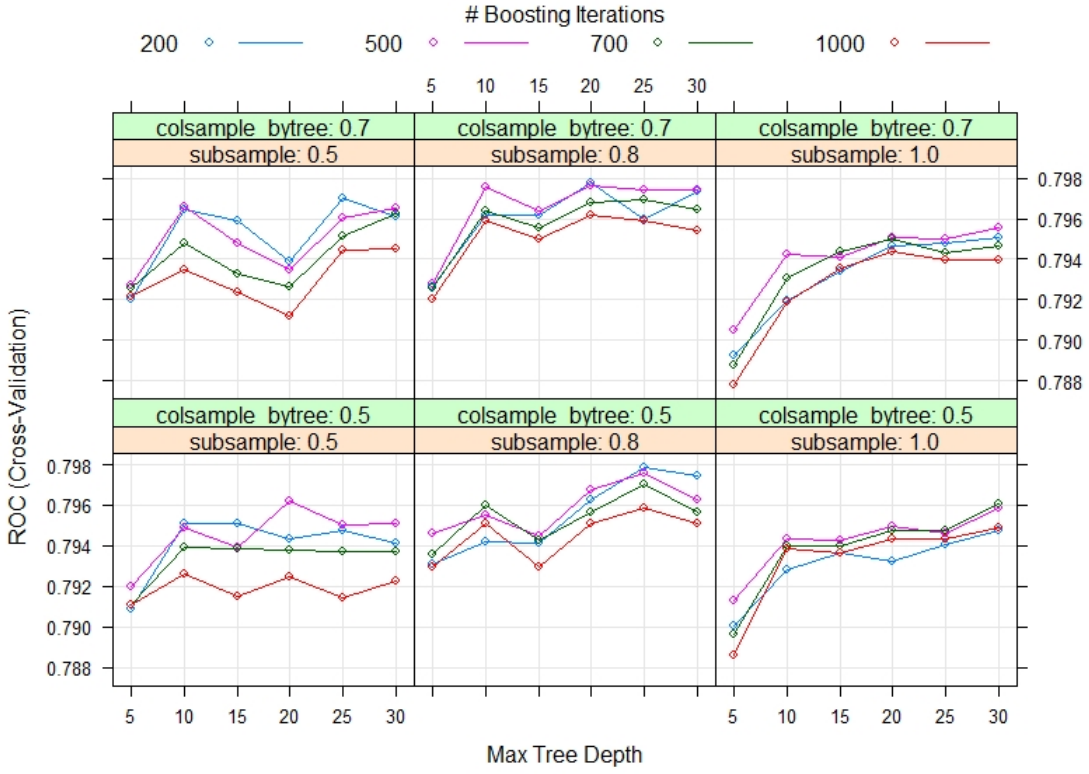



FIGURE 5.8 – Tuning des paramètres *XGBoost* sur la base *vol*

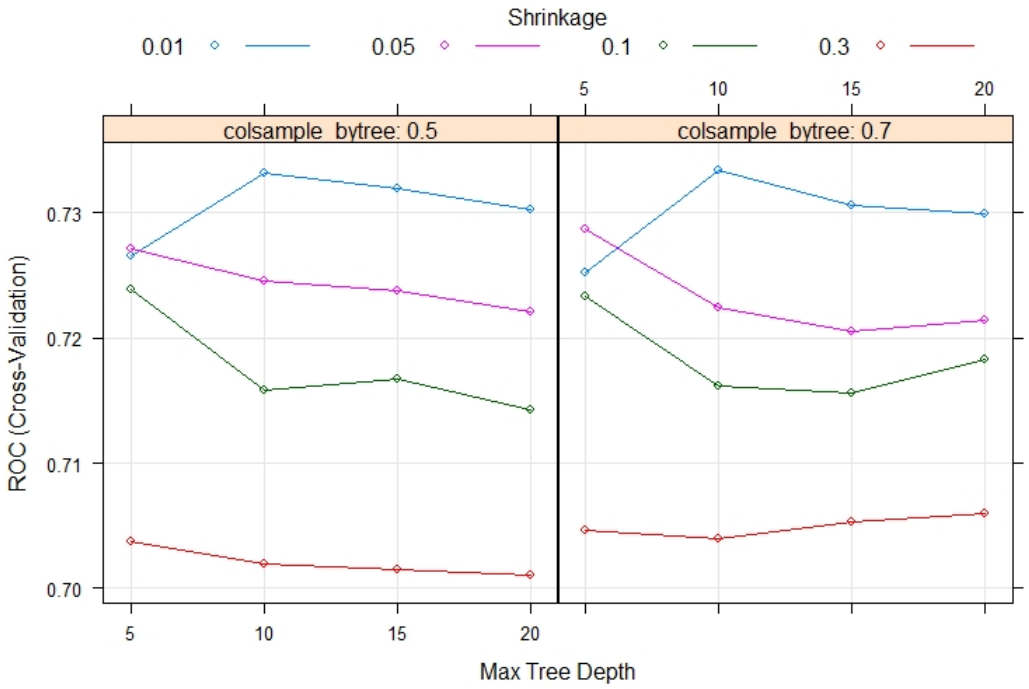



FIGURE 5.9 – Tuning des paramètres XGBoost sur la base *circulation*

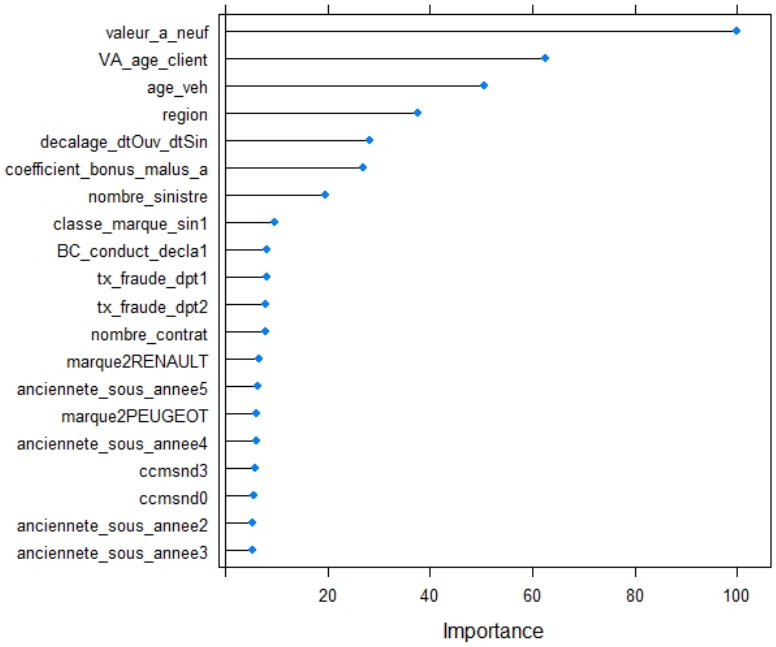



FIGURE 5.10 – Importance des variables par XGBoost pour la base *incendie*

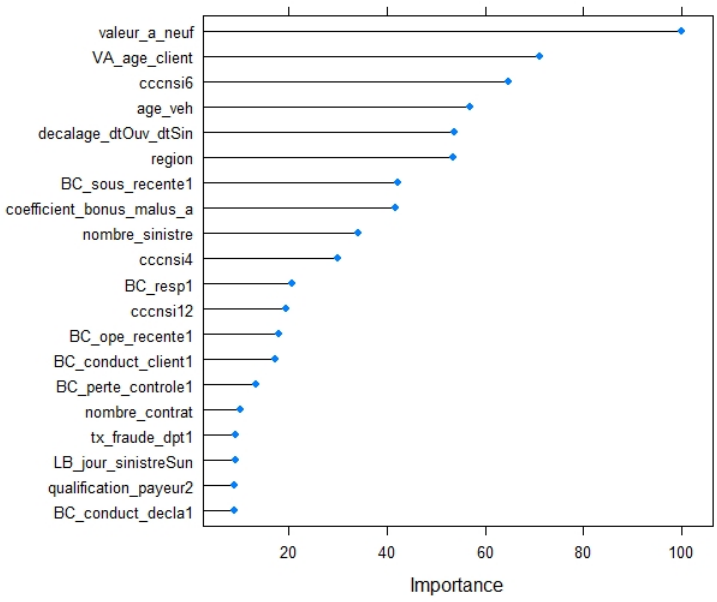



FIGURE 5.11 – Importance des variables par XGBoost pour la base *circulation*

## Annexe 6 : Liste des variables

La liste des variables n'est pas exhaustive mais comporte les variables les importantes dans la modélisation.

| Variable | Signification | Type |
|---|---|---|
| BC_ope | Opération récente | Catégorielle |
| BC_sous_recente | Souscription récente | Catégorielle |
| sinistre_recent | Sinistralité récente | Catégorielle |
| assistance_recente | Assistance récente | Catégorielle |
| qualification_payeur | Qualité de paiement (bon payeur ou non) | Catégorielle |
| VA_age_client | Age de l'assuré | Continue |
| Age_veh | Ancienneté du véhicule | Continue |
| Valeur_a_neuf | Valeur du véhicule | Continue |
| Cccnsi | Circonstances du sinistres | Catégorielle |
| Ccmsnd | Motif de déplacement | Catégorielle |
| Precision_vol | Complément circonstances vol | Catégorielle |
| BC_resp | Taux de responsabilité | Catégorielle |
| BC_rapport de police | Rapport de police ou PV | Catégorielle |
| Nombre_sinistre | Nombre de sinistre | Continue |
| Nombre_contrat | Nombre de contrats détenus | Continue |
| BC_jour_ferie | Jour de sinistre correspond à un jour férié | Catégorielle |
| anciennete_sous_annee | Ancienneté du contrat | Catégorielle |
| Classe_tx_vol | Classement du taux de vol par département | Catégorielle |
| Région | Région du sinistre | Catégorielle |
| Nombre_panne | Nombre de pannes | Continue |
| Prcnt_panne | Pourcentage de pannes non prises en charge | Continue |
| LB_jour_sinistre | Nom du jour de sinistre | Catégorielle |
| Fraude2 | Fraude | Catégorielle |