Mémoire présenté le :

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

Par : Valentin LORANGE

Titre **Etude et prédiction des sinistres graves en assurance Multirisque Commerce à l'aide du Machine Learning**

Confidentialité : ☐ NON ☑OUI (Durée : ☐ 1 an ☑2 ans)

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise : Generali

Nom :

Signature :

Directeur de mémoire en entreprise :

Nom : Ubezzi Robin

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

Mémoire Actuaire

Étude et prédiction des sinistres graves en assurance Multirisque Commerce avec du Machine Learning

réalisé chez Generali France

LORANGE Valentin

Encadrement
Ubezzi Robin - Tuteur en entreprise
Planchet Frédéric - Tuteur académique

ISFA
13/02/2022

Contexte de l'étude et objectifs

Ce mémoire a pour objectif d'étudier les sinistres graves du portefeuille Multirisque Commerce de Generali France survenus ces dernières années dans l'intention de prévenir ce type de sinistre dès la fin d'année 2021.

Cette étude s'insère ainsi dans un projet plus large réalisé au sein de Generali. Nommé LCAP (Large Claims Alert Project), ce projet a pour objectif de "dérisquer" les portefeuilles Multirisque Commerce et Risques Industriels à la suite d'une augmentation constatée de l'importance des sinistres graves à la fois sur le plan de la fréquence et sur le plan du coût moyen. Il s'agit d'identifier les principaux facteurs de risques de la sinistralité grave.

Résumé

Mots clés : Assurance multirisque commerce, sinistre grave, modèle de propension, théorie des valeurs extrêmes, CART, Random Forest, Gradient Boosting, SHAP, AUC, Classification binaire

Dans ce mémoire, on modélise à des fins prédictives et/ou de prévention la sinistralité grave en assurance Multirisque Commerce. Le modèle ne s'intéresse qu'au fait qu'un sinistre soit un sinistre grave ou non : c'est un modèle de propension. En d'autres termes, on s'intéresse à estimer la probabilité qu'un sinistre survenu conditionnellement au risque associé soit grave.

Après construction de la base obtenue en regroupant des données sinistres, des données contrats, des données risque et des données externes, un retravail et une analyse des données sont effectués. Les données manquantes sont en partie gérées par une méthode d'implantation multiple séquentielle. Les modalités trop nombreuses sont gérées avec un algorithme des plus proches voisins.

Un sinistre grave est défini par l'intermédiaire d'un seuil fixe qui, une fois dépassé, définit ce sinistre comme grave. La définition de ce seuil est habituellement arbitraire. La performance et la qualité de la modélisation sont extrêmement liées à la valeur de ce seuil. L'utilisation d'outils mathématiques provenant de la Théorie des Valeurs Extrêmes est alors nécessaire afin de déterminer ce seuil. L'utilisation d'un Q-Q plot confirme le caractère extrême des sinistres MRC. La combinaison de plusieurs méthodes (Mean Excess Function, estimateurs de ξ, Gerstengarbe plot) nous amène à sélectionner deux seuils. Un premier seuil utilisant la charge du sinistre (100 000€) et un deuxième utilisant un taux de destruction (12%).

Lors de l'élaboration du modèle de propension, on est amené à utiliser des algorithmes de classification supervisée binaire issus du Machine Learning. Les deux méthodes retenues sont des méthodes basées sur des arbres CART :

Random Forest et Gradient Boosting. On optimise les modèles en se basant principalement sur l'AUC et avec une méthode de type Grid Search et en utilisant la validation croisée. On étudie l'impact du seuil de classification retenue en analysant des matrices de confusions. Le meilleur algorithme est l'algorithme de type Gradient Boosting puisqu'il fournit de meilleurs résultats pour les deux seuils utilisés en matière de qualité de prédiction.

Une fois la modélisation du score de propension, les principaux facteurs de risques sont identifiés avec notamment l'utilisation du coefficient SHAP afin de répondre à la problématique initiale de "dérisquer" le portefeuille.

Summary

Key words : Commercial Multi-risk Insurance, large claims, propensity model, extreme value theory, CART, Random Forest, Gradient Boosting, SHAP, AUC, Binary classification

In this thesis, we try to model for predictive and/or preventive purposes the large claims in commercial multi-risk insurance. The model is only interested in the fact that a claim is a large claim or not : it is a propensity model. In other words, we are interested in estimating the probability that a claim occurring is a large claim conditionally to the associated risk.

After the construction of the database obtained by grouping claims data, contract data, risk data and external data, a reworking and an analysis of the data are carried out. Missing data are partly managed by a sequential multiple implantation method. The too numerous modalities are managed with a nearest neighbor algorithm.

A large claim is defined by a fixed threshold which, once exceeded, defines the claim as large. The definition of this threshold is usually arbitrary. The performance and quality of the modeling is highly dependent on the value of this threshold. The use of mathematical tools from the Extreme Value Theory is then necessary to determine this threshold. The use of a Q-Q plot confirms the extreme character of the Commercial Multi-risk claims. The combination of several methods (Mean Excess Function, ξ estimators, Gerstengarbe plot) leads us to select two thresholds. A first threshold using the claim load (100 000€) and a second using a destruction rate (12%).

During the development of the propensity model, we use binary supervised classification algorithms from Machine Learning. The two methods chosen are based on CART trees : Random Forest and Gradient Boosting. The models are optimized mainly based on the AUC and with a Grid Search method and using cross validation. We study the impact of the chosen clas-

sification threshold by analyzing confusion matrices. The best algorithm is the Gradient Boosting algorithm since it provides better results for the two thresholds used in terms of prediction quality.

Once the propensity score has been modeled, the main risk factors are identified, notably by using the SHAP coefficient in order to respond to the initial problem of "derisking" the portfolio.

Remerciements

Je tiens tout d'abord à remercier toutes les équipes de la Direction Technique Non Vie de Generali France de m'avoir accueilli et de m'avoir offert une expérience instructive.

Je tiens en particulier à remercier toute l'équipe Indemnisation dont je fais partie. Notamment Vincent Cepa pour son aide et bien sûr mon tuteur Robin Ubezzi pour son accompagnement, ses conseils et pour s'être tenu disponible et m'avoir fait confiance tout le long de mon travail.

Je remercie également l'ensemble de l'équipe LCAP composée de Hélène Deboudt, Hassan Sedraoui, Massile Mourah, Ambre Le Stum, Ismail Hammounou, Pierre Desmet et Yoann Gouyen.

Table des matières

Abréviations

AUC = Area Under Curve
BDG = Bris de glace
BDM = Bris de machine
CA = Chiffre d'affaires
CART = Classification And Regression Tree
DDE = Dégâts des eaux
DEL = Dégâts électriques
FFB = Fédération Française du Bâtiment
GEV = Generalized Extreme Value
GLM = Generalized Linear Model
GPD = Generalized Pareto Distribution
MEF = Mean Excess Function
MRC = Multirisque Commerce
MRH = Multirisque Habitation
MRI = Multirisque Immeuble
PE = Perte d'exploitation
RC = Responsabilité Civile
TCL = Théorème Central Limite
TD = Taux de destruction

Generali et l'assurance MRC

Dans cette partie, on commence par présenter l'assurance MRC puis on donne quelques précisions sur le cas particulier de Generali. Enfin, on expose le mode de gestion des sinistres graves chez Generali.

1.1 Présentation de l'assurance MRC

Le produit Multirisque Commerce (MRC) est une assurance à destination de professionnels comme les artisans, les commerçants, les professions libérales, les professionnels du bâtiment, mais aussi les exploitants agricoles ou les autoentrepreneurs. Lors de la souscription, le client décide ainsi d'assurer un ou plusieurs bâtiments/commerces et des biens (machines, équipements, stocks ...). Un capital doit être convenu pour estimer la valeur de ces biens. L'assurance MRC couvre des dommages "involontaires", elle ne couvre donc pas contre la vétusté.

Elle contient de nombreuses garanties, voici les plus importantes :

— La garantie Responsabilité Civile qui protège contre tous types de dommages (matériels, immatériels, corporels) causés à des clients, partenaires ou fournisseurs lors de l'activité de l'entreprise.

— La garantie Vol / Vandalismes qui protège contre la disparition ou la détérioration de biens suite à un acte criminel.

— La garantie Dégâts des eaux qui couvre contre des dégâts causés par une fuite d'eau, une rupture de canalisation, des infiltrations d'eau sous le toit ...

— La garantie Bris de machine qui permet de couvrir des dégâts causés à tout appareil de production (machine, chaudière, engin de manutention ...)

— La garantie Bris de glace qui concerne tout ce qui est fenêtre, vitrine, miroirs ...

— La garantie Dégâts électriques qui couvre les dégâts causés à tous les équipements électroniques suite à un problème électrique (court-

circuit, surtension ...).
— La garantie Perte d'exploitation qui permet de faire face à une perte de revenue temporaire à la suite d'un sinistre
— La garantie catastrophes naturelles qui concerne les dommages suite à des événements comme les inondations, la sécheresse ...
— La garantie incendie qui intervient en cas d'incendie.

Dans le compartiment des assurances de dommages aux biens[1], la MRC est la troisième plus importante en termes de cotisations (8.2 milliards d'euros en 2020) derrière l'assurance Auto et l'assurance MRH. Elle croît de 2 à 3 % par an et un peu plus de deux millions d'entreprises sont assurées. Elle représente 11% des sinistres survenus en 2020. Et il y a eu à peu près 6 milliards d'euros de prestation en 2020. Le ratio combiné après réassurance est de l'ordre de 98%.

1.2 La MRC chez Generali

Chez Generali France, l'offre MRC est distribuée sous quatre formes selon le distributeur et le type de commerce. Pour chacun de ces quatre produits, il y a trois formules. Les formules 1 et 2 qui définissent des niveaux de garantie et de franchise fixes et une formule libre avec des garanties et des franchises aux choix.

La prime est calculée de manière traditionnelle en calculant une prime pour chaque garantie. Pour certaines garanties comme la perte d'exploitation, la prime est calculée à la maille contrat. Pour d'autres garanties comme l'incendie, la prime est calculée pour chaque site. Mais la plupart du temps, il n'y a qu'un seul site.

$$P = \sum_{i=1}^{Nb\_Sites} P_i$$

$$P_i = \sum_{k=1}^{Nb\_Garanties} P_{ik}$$

Chaque prime P_{ik} est obtenue par un modèle GLM multiplicatif. Une prime de référence est calculée par un modèle coût-moyen / fréquence pour les si-

1. Fédération Française de l'Assurance : https://www.ffa-assurance.fr/etudes-et-chiffres-cles (accès le 28/07/2021)

nistres attritionnels. Ensuite selon les différents facteurs de risques, on multiplie cette prime de référence par un nombre associé à chaque risque. Si le facteur est supérieur à 1 alors c'est que le contrat qu'on est en train de tarifer est plus risqué que la référence pour le risque considéré. Inversement, il est moins risqué si le facteur est inférieur à 1.

$$\tilde{P}_{ik} = \rho_k \cdot \prod_{r=1}^{Nb\_Risques} \rho_{kri}$$

ρ_k est la prime de référence pour la garantie k. ρ_{kri} est le facteur de risque associé à la garantie k et au risque r du site i. Certains risques sont liés à la localisation de l'entreprise, d'autres sont liés au type de commerce, d'autres sont liés à l'entreprise en elle même.
On arrive ainsi à une prime pour la garantie considérée. Cette prime $\tilde{P}_{ik}$ est ensuite ajustée selon certaines spécificités du contrat (la franchise par exemple) ou selon certaines caractéristiques non prises en compte dans le modèle (présence d'extincteur, ramonage ...). Ces ajustements ne sont pas issus de modèles spécifiques, mais sont plus "à dire d'expert". Une fois cela fait, un taux de surcrête est appliqué. En effet, la prime ne concerne que les sinistres attritionnels : on a exclu les sinistres graves. Ce taux de surcrête ne dépend que de la garantie : elle ne dépend pas du niveau de risque spécifique au contrat que l'on cherche à tarifer. Le risque grave est mutualisé entre tous les assurés : il y a "solidarité".

$$P_{ik} = (1 + r) \cdot \Phi(\tilde{P}_{ik})$$

Ce taux de surcrête r est calculé à partir de la base initiale au moment où on retire des sinistres graves. Avec W le coût d'un sinistre :

$$r = \frac{\sum_{s=1}^{Nb\_Sinistres} W_s \cdot \mathbb{1}_{W_s > seuil}}{\sum_{s=1}^{Nb\_Sinistres} W_s \cdot \mathbb{1}_{W_s < seuil}}$$

On obtient de cette manière la prime associée à une garantie et à un site. On a ainsi obtenu la prime pure P associée au contrat. Cette prime est ensuite revalorisée en prenant en compte le taux de commission, les différents frais, le taux de réassurance, le coût du capital, les taxes ... On obtient ainsi la prime proposée au client.

Chez Generali, il y a environ 20000 sinistres par an pour un coût moyen d'un peu plus de 4000€. Les trois principales garanties en ce qui concerne les charges sont la garantie incendie (29%), la garantie dégâts des eaux (17%) et la garantie perte d'exploitation (16%).

1.3 Gestion des sinistres graves

Un sinistre est dit grave chez Generali quand sa charge est supérieure à 150000€. Ainsi, lors de l'enregistrement d'un nouveau sinistre, une estimation du montant du sinistre est effectuée pour le provisionnement. Cette estimation évolue au fur et à mesure que le dossier évolue lui aussi. Si lors de l'évolution du dossier, le montant estimé se révèle être supérieur à 150000€ alors la gestion du dossier est confiée à un organisme spécifique du service indemnisation et une fiche dite "fiche grave" est alors créée. La procédure de gestion du sinistre est ainsi différente et en particulier il y aura au moins un rapport d'expertise effectué.

Base de données

Dans cette partie, on commence par présenter les différentes données à notre disposition selon les différentes sources. Ensuite, on réalise une brève analyse descriptive des données. Puis on s'intéresse au problème des données manquantes. Enfin, on précise comment ont été gérées les variables catégorielles avec un trop grand nombre de modalités.

2.4 Composition de la base

Le but de cette étude étant d'avoir une qualité prédictive, il fallait le plus de données possible. On va donc considérer l'ensemble des sinistres survenus entre 2013 et 2020. Le choix de 2013 était le meilleur compromis entre nombre de sinistres (qu'on cherche le plus grand possible) et complétude des informations (plus on remonte loin dans le passé, moins on a d'informations).

On retire d'abord les sinistres sans suite ou annulés. On a aussi fait le choix de retirer les sinistres RC et climatiques. Les sinistres climatiques sont des événements dont le risque ne dépend pas du comportement humain : c'est purement un risque lié au climat (une section chez Generali nommé Climate Lab est spécialisée dans la modélisation de ce risque). Les sinistres RC sont enlevés, car ils concernent des sinistres de type corporel et notre étude vise à regarder les dommages matériels. On a aussi retiré les groupements de notre portefeuille MRC étant donné que la nature du risque est totalement différente : il faudrait faire une étude séparée sur ce type de risque.

Une fois ce périmètre fixé, il a fallu aller chercher le plus de données possibles : certaines données sont liées aux sinistres, certaines sont des données contrats, d'autres des données risques et enfin il y a aussi quelques données externes. Une ligne i de notre base correspond à un sinistre et les colonnes j correspondent aux différentes informations que l'on a récupérées.

2.4.1 Données Sinistres

On a récupéré beaucoup de données sur les sinistres à étudier afin de mieux connaitre ce qu'on allait modéliser. Mais la plupart de ces données ne seront pas utilisées dans la modélisation étant donné que l'on cherche prioritairement un modèle prédictif. Toutes les variables a posteriori sont donc laissées de côté.

On a décidé d'utiliser les données sinistres d'une nouvelle manière : en utilisant la sinistralité passée. C'est-à-dire qu'on a récupéré des informations sur les sinistres survenus dans les trois années précédentes. Ainsi si un sinistre est survenu en 2015 alors on a regardé le passé de cet assuré entre 2012 et 2014. Les sinistres que l'on regarde ne concernent pas que les sinistres liés au contrat MRC, mais l'ensemble des contrats du client s'il est multiéquipé : on peut avoir des contrats Auto/Flottes ou bien du MRH par exemple. On peut remarquer que l'on n'a pas toujours cette information, car il se peut que le client n'ait pas d'autres contrats chez Generali ou alors qu'il soit un nouveau client. Les sinistres passés ne sont pas utilisés lors de la tarification en MRC (contrairement à l'Auto) car ces données sont uniquement déclaratives, elles ne sont donc pas fiables. Mais dans le cadre de notre étude, on peut utiliser les informations sur les contrats que l'on avait déjà en portefeuille.

On a aussi décidé de regrouper en une seule ligne les sinistres connexes. En effet, la manière de gérer les sinistres fait que si pour un événement (un incendie un exemple), plusieurs garanties entrent en jeu (Incendie et Perte d'exploitation par exemple) alors plusieurs sinistres sont ouverts. Dans notre exemple, les deux sinistres Incendies et PE ne sont considérés que comme un seul sinistre. Si la garantie Incendie a coûté 100 000€ et la garantie PE 50 000€, on aura dans notre base un seul sinistre de charge 150 000€. Par contre, s'il y a aussi eu un accident corporel alors la garantie RC associée ne sera pas prise en compte comme expliqué précédemment.

Les charges de sinistres ont été revalorisées en fonction de leur date de survenance avec une mise en "As if". Cela permet de prendre en compte l'inflation. Cela est nécessaire, car on travaille sur la gravité des sinistres et donc sur le montant. On a utilisé pour cela l'indice FFB[2] du coût de la construction : il est fourni par la Fédération Française du Bâtiment de manière trimestrielle. Il permet de suivre l'évolution du coût de la construction

2. Fédération Française du Bâtiment : https://www.ffbatiment.fr/federation-francaise-du-batiment/le-batiment-et-vous/en_chiffres/indices-index/Chiffres_Index_FFB_Construction.html (accès le 04/07/2021)

d'un immeuble et il est souvent utilisé pour l'indexation des polices en MRC, MRI et MRH.

Exemple :
L'indice FFB vaut 1022,3 au T1 2021. Prenons un sinistre survenu au T2 2017. L'indice valait 960.1. On revalorise la charge de ce sinistre en la multipliant par $\frac{1022.3}{960.1} = 1.065$.

Variable	Définition	Remarques
S_CLE	Clé du sinistre obtenue après avoir rassemblé les sinistres connexes	unique (permet d'identifier la ligne)
$S\_DATE\_SURVENANCE$	Date de survenance du sinistre	vaut -1 quand l'exposition est nulle
S_CHARGE_BRUTE	Charge du sinistre (somme cumulée si sinistre connexe)	variable à modéliser
$S\_GARANTIE$	Nature du sinistre (nature du plus gros sinistre si connexe)	Variable non utilisée dans la modélisation
$S\_INCENDIES$	Fréquence antérieure des sinistres Incendie	vaut -1 quand l'exposition est nulle
S_DEL	Fréquence antérieure des sinistres Dégâts électriques	vaut -1 quand l'exposition est nulle
S_BDG	Fréquence antérieure des sinistres Bris de Glace	vaut -1 quand l'exposition est nulle
S_VOL	Fréquence antérieure des sinistres Vol/Vandalisme	vaut -1 quand l'exposition est nulle
S_DDE	Fréquence antérieure des sinistres Dégâts des eaux	vaut -1 quand l'exposition est nulle
S_FREQ_RI	Fréquence antérieure en Risque industrielle	vaut -1 quand l'exposition est nulle
S_FREQ_MRC	Fréquence antérieure en MRC (autres que ceux listés précédemment)	vaut -1 quand l'exposition est nulle
$S\_FREQ\_FLOTTES$	Fréquence antérieure en Flottes	vaut -1 quand l'exposition est nulle
S_FREQ_AUTO	Fréquence antérieure en Auto (hors flottes)	vaut -1 quand l'exposition est nulle
$S\_FREQ\_TRANSPORT$	Fréquence antérieure en Transport	vaut -1 quand l'exposition est nulle
S_FREQ_MRH	Fréquence antérieure en MRH	vaut -1 quand l'exposition est nulle
S_FREQ_MRI	Fréquence antérieure en MRI	vaut -1 quand l'exposition est nulle

TABLE 2.1 – Définitions des principales variables Sinistre

2.4.2 Données Contrats

On a récupéré dans la base Contrats les différentes caractéristiques de l'assuré (identité, âge, sexe, informations sur la facturation ...), les informations sur la police (garanties, date d'effet, date de résiliation ...) et les informations sur le commerce.

Variable	Définition	Remarques
$C\_TERCODC$	Code territoire (permet de différencier Paris/Province/Outre mer/Pays étranger	variable catégorielle
$C\_POLPTAM$	Prime annuelle	
C_CODE_INT	Code intermédiaire (identifie l'intermédiaire)	variable catégorielle
$C\_CODE\_POLE\_RESEAU$	Code du Réseau (identifie le réseau)	variable catégorielle
$C\_ENGAGEMENT$	Engagement de la police	
$C\_FORMULSI1$	Formule choisie pour la police	3 modalités
$C\_NBSITES$	Nombre de sites	
C_CLASSE_INC	Classe tarifaire Incendies	
C_CLASSE_DDE	Classe tarifaire Dégâts des eaux	
C_CLASSE_PE	Classe tarifaire Perte d'exploitation	
C_CLASSE_BDM	Classe tarifaire Bris de machines	
$C\_QUALITE$	Qualité de l'occupant	propriétaire,locataire,copropriétaire
$C\_FORMEJURI$	Forme juridique de l'entreprise	plusieurs dizaines de modalités

TABLE 2.2 – Définitions des principales variables Contrat

Une classe tarifaire est un regroupement permettant de synthétiser le risque : les entreprises d'une même classe ont un risque homogène concernant la garantie considérée. Les éléments pris en compte sont des informations sur l'entreprise et son activité. Les éléments géographiques sont utilisés dans le zonier. Par exemple, les artisans ayant des activités industrielles ou chimiques (comme un verrier) sont plus à risque concernant les incendies qu'un cabinet d'avocat.

La formule détermine le niveau des garanties pour le contrat. Certaines garanties comme la garantie Incendies sont obligatoires, d'autres comme la garantie Bris de Machines ne le sont pas. Il faut alors déterminer si la garantie

est souscrite ou non et le cas échéant, son niveau d'indemnisation (entre 0% et 100%).

2.4.3 Données Risque

Les données risques sont sans doute les données les plus importantes de la base. La plupart de ces données sont déjà utilisées dans les modèles de tarification. Voici les principales :

Variable	Définition	Remarques
R_DEP	Numéro de département	
$R\_SURFACE$	Surface du commerce assuré	
R_CODE_NAF	Code de Nomenclature d'activité française	plusieurs dizaines de modalités
$R\_EFFECTIF$	Effectif du commerce	
R_CA	Chiffre d'affaires du commerce	
R_CAPINC	Capital Incendie	
R_CAPDDE	Capital dégâts des eaux	
R_CAPVOL	Capital Vol	
R_CAPBDM	Capital Bris de Machine	
$R\_DTCREAT\_ENTR$	Date de création de l'entreprise	beaucoup de vides
R_GROUP_ACT	Code qui caractérise l'activité (de manière large) de l'entreprise	4 modalités
$R\_EXCLUBAT\_RISQUE\_LOC$	Exclusion risque locatif	vaut 0 ou 1
$R\_ASS\_CPTE\_PROPR$	Assurance pour compte du propriétaire	vaut 0 ou 1
$R\_LIB\_PROTECT\_VOL$	Indique le type (le cas échéant) de protection contre les vols	
$R\_MATERIAUX\_DURS$	Indique si le bâtiment est fabriqué à partir de matériaux durs à plus de 75%	vaut 0 ou 1
$R\_INSTA\_ELEC\_VERIF$	indique (le cas échéant) la fréquence des vérifications électrique	
$R\_VERIF\_EXTINCTEUR$	indique (le cas échéant) la fréquence des vérifications des extincteurs	
R_ZONE_INC	Zone tarifaire Incendies	
R_ZONE_BDM	Zone tarifaire Bris de machines	
$R\_VETUSTE$	Informations sur la vétusté	

TABLE 2.3 – Définitions des principales variables Risque

2.4.4 Données Externes

On a aussi réussi à ramener quelques données externes. La première donnée intéressante est la distance aux pompiers. En effet, dans la partie suivante on constate que les incendies représentent une grande partie des sinistres graves. Ainsi la distance entre le commerce et la station de pompier la plus

proche peut être un bon indicateur.

Une deuxième donnée externe concerne la santé financière de l'entreprise sur les années qui précèdent. Il est possible de penser qu'une entreprise en bonne santé d'un point de vue financier aura les moyens d'appliquer les différentes démarches recommandées pour limiter le risque de sinistre (formation du personnel, mise en place de dispositifs antivol de bonne qualité, vérifications fréquentes du réseau électrique ...). Au contraire, une entreprise qui a des difficultés financières pourrait avoir tendance à négliger tout ce qui est prévention. On s'attend ainsi à ce que la santé financière soit un bon indicateur. Les notes utilisées sont des notes allant de 0 à 7. 7 correspond à un excellent état financier alors que 0 correspond à un état de faillite. On a décider de prendre deux notes par sinistre : une pour l'année du sinistre et l'autre pour l'année précédente. Par exemple, si le sinistre a eu lieu en 2015, alors on a les notes au 01/01/2015 et au 01/01/2014.

On a aussi réussi à récupérer la date de création de l'entreprise (dans 92% des cas) et l'ancienneté du bâtiment (dans seulement 30% des cas).

Enfin, la dernière information intéressante est le prix du m^2 qui a été calculé par code postal et par année. Mais cette information n'est pas toujours disponible.

Ces données sont nouvelles dans la modélisation de la sinistralité en MRC. Un des objectifs secondaires de ce projet est donc de donner un premier aperçu du pouvoir explicatif de ces différentes variables.

2.5 Analyse descriptive

2.5.1 Analyse des sinistres

Il y a 99552 sinistres, dont 500 avec une charge supérieure à 100000€ et 987 avec un taux de destruction supérieur à 12%. Cette définition pour les sinistres graves est détaillée ultérieurement.

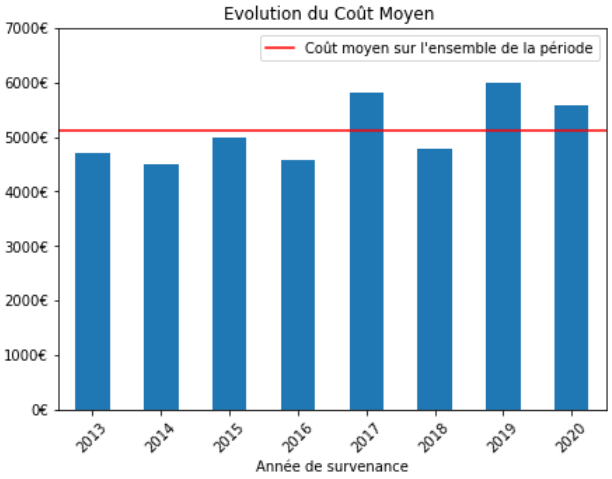

FIGURE 2.1 – Évolution du coût moyen

Commençons par regarder l'évolution du coût moyen des sinistres. Ce coût moyen est à peu près constant (écart-type de 570€ pour une moyenne de 5116€) mais on observe quand même une légère tendance à la hausse. Cette hausse peut provenir soit de la mise en "as if" qui ne correspondrait pas à la réalité soit d'une réelle augmentation du coût d'un sinistre.

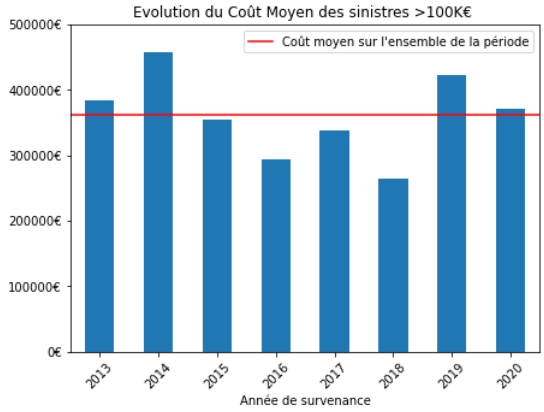

FIGURE 2.2 – Évolution du coût moyen pour les sinistres supérieurs à 100 000 €

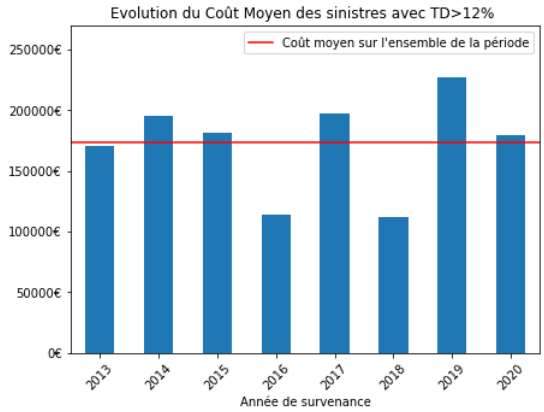

FIGURE 2.3 – Évolution du coût moyen pour les sinistres avec un TD>12%

Concernant le coût moyen des sinistres graves, il y a une plus grande volatilité, ce qui est normale étant donné le caractère beaucoup plus aléatoire et les faibles effectifs. La seconde chose que l'on remarque est la différence de coût moyen entre les deux types de graves. En effet le coût moyen des

graves par charge brute est à peu près deux fois supérieur au coût moyen des graves par taux de destruction. Ce qui veut dire que l'approche par taux de destruction ramène beaucoup de "petits" sinistres.

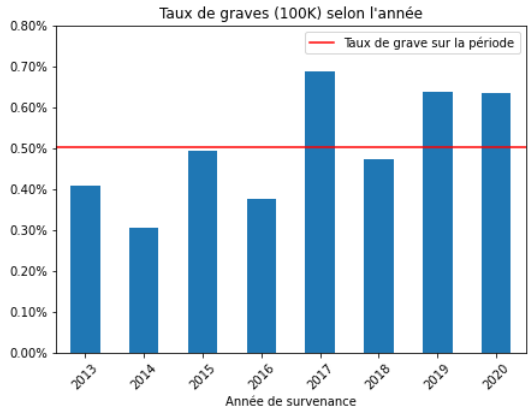

FIGURE 2.4 – Évolution du taux de graves (>100k€)

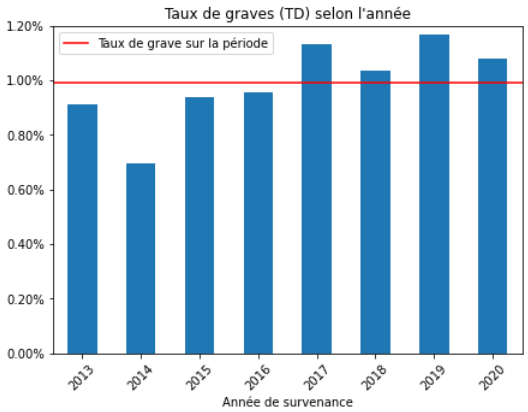

FIGURE 2.5 – Évolution du taux de graves (>12%)

Le taux de grave (en nombre) est clairement orienté à la hausse, quelle que soit la méthode. Ce résultat était attendu : c'est pour cela que cette étude sur les graves en MRC a été réalisée.

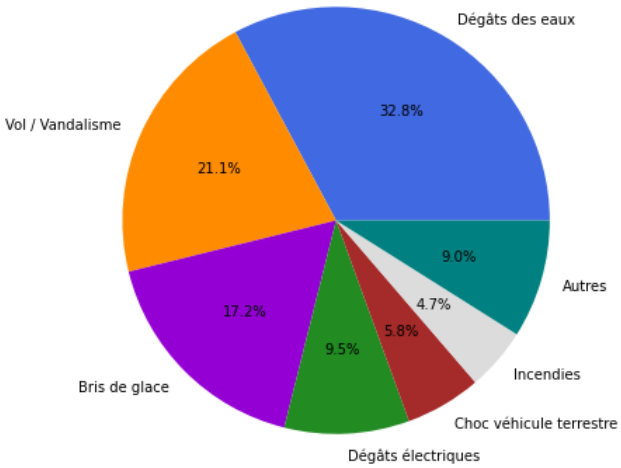

FIGURE 2.6 – Répartition (en nombre) par nature de sinistre

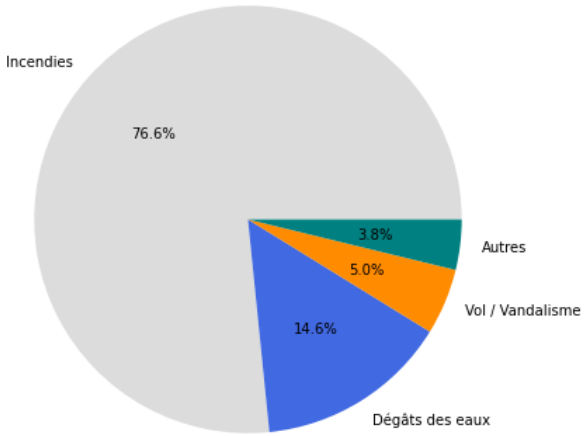

FIGURE 2.7 – Répartition (en nombre) par nature de sinistre des graves (>100k€)

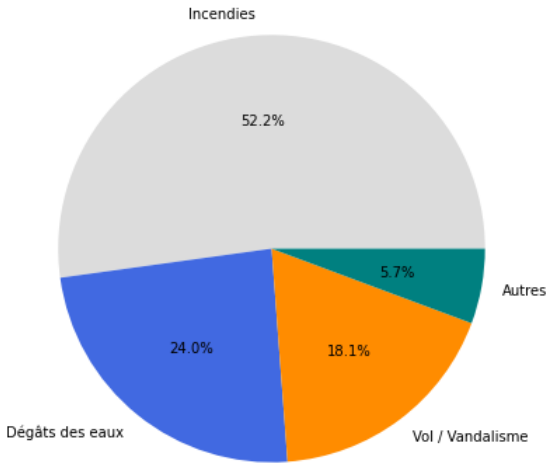

FIGURE 2.8 – Répartition (en nombre) par nature de sinistre des graves (>12%)

L'étude de la nature des sinistres montre la prédominance des incendies dans les graves. Ils représentent moins de 5% des sinistres, mais la grande majorité des graves sont des incendies. On remarque aussi la présence en plus forte proportion des sinistres dégâts des eaux et vol/vandalisme dans l'approche par taux de destruction. Cette méthode semble ramener des sinistres de natures différentes. On peut en profiter pour ajouter le fait que seulement 1.5% des sinistres ont présence de PE alors que pour les sinistres supérieurs à 100k€, il y en a 61.6% et pour les sinistres avec un taux de destruction supérieur à 12% il y en a 34.1%. La PE est donc quelque chose qui caractérise un sinistre grave : ce qui est facilement explicable puisqu'un "gros" sinistre implique en général un arrêt conséquent de l'activité de l'entreprise (incendie, dégâts des eaux).

	Charge<100k€	Charge>100k€	Total
TD<12%	98422	143	98565
TD>12%	630	357	987
Total	99052	500	99552

TABLE 2.4 – Tableau croisé selon le type de grave

De manière graphique :

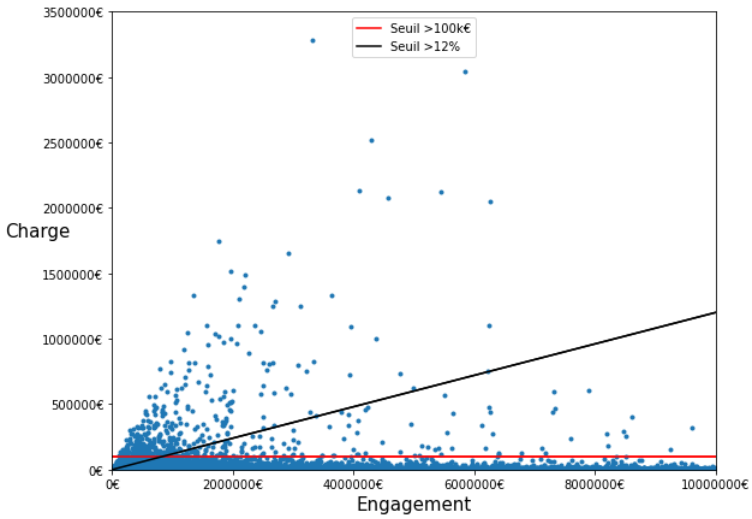

FIGURE 2.9 – Échelle linéaire

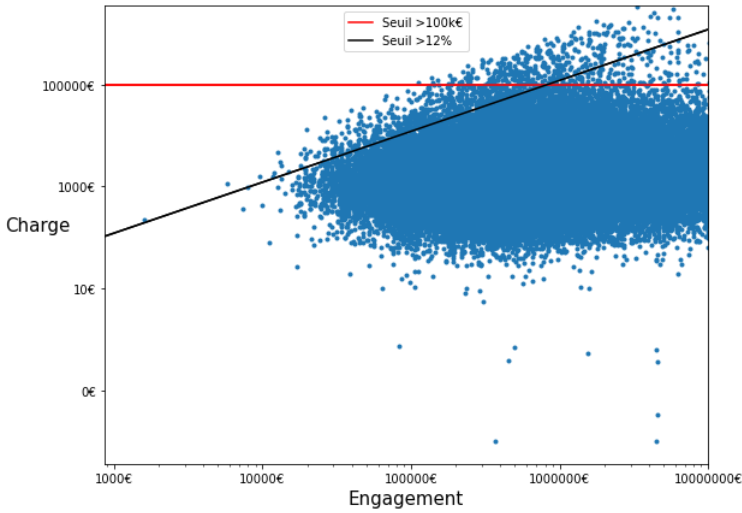

FIGURE 2.10 – Échelle log

Encore une fois, le tableau croisé montre bien la différence entre les deux approches de la gravité d'un sinistre. L'approche par taux de destruction permet de ramener des sinistres avec une charge "faible", mais qui ont un engagement encore plus faible et élimine certains sinistres avec une charge "élevée" mais qui est expliqué par un engagement important. Ceci est bien illustré sur le graphe de gauche, de nombreux points sont au-dessus de la droite rouge, mais bien en dessous de la droite noire. Sur le graphe de droite, on remarque beaucoup de sinistres qui ont couté quelques dizaines de milliers d'euros, mais qui correspondaient à des petits engagements, ils sont graves au sens du taux de destruction.

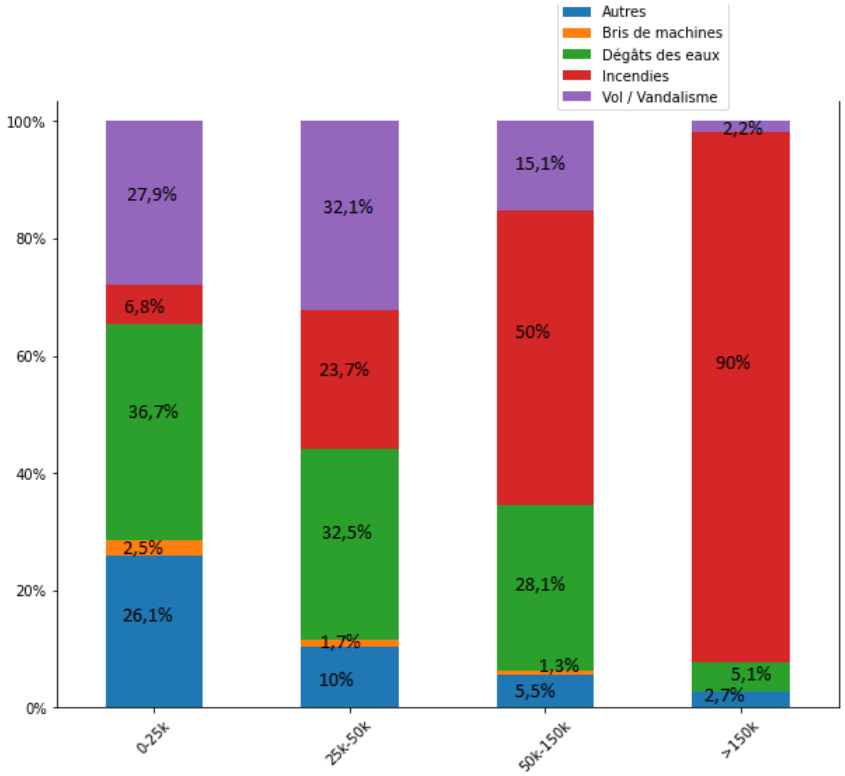

FIGURE 2.11 – Répartition (en charge) par nature de sinistre selon la tranche

L'importance des sinistres (en charge) selon la tranche montre la montée progressive des incendies dans le poids total quand la tranche est de plus en plus élevée. Pour résumé, on peut dire que les sinistres graves en charge sont principalement des incendies et que l'approche par taux de destruction met en évidence en plus des sinistres Vol/Vandalisme et Dégâts des eaux.

2.5.2 Analyse du portefeuille sinistré

On va, dans cette partie, regarder quelques indicateurs pour en apprendre plus sur le portefeuille sinistré. Ceux ne sont pas des statistiques sur l'ensemble du portefeuille MRC de Generali mais bel et bien sur les contrats ayant eu un sinistre que l'on étudie dans cette étude. En particulier, une seule entreprise ayant eu deux sinistres différents comptera pour deux.En fonction du profil de sinistralité, il se peut que les informations soient donc différentes du portefeuille dans sa globalité. On ne s'intéresse pas au portefeuille entier, car la modélisation porte sur un modèle de propension : on ne s'intéresse qu'aux sinistres.

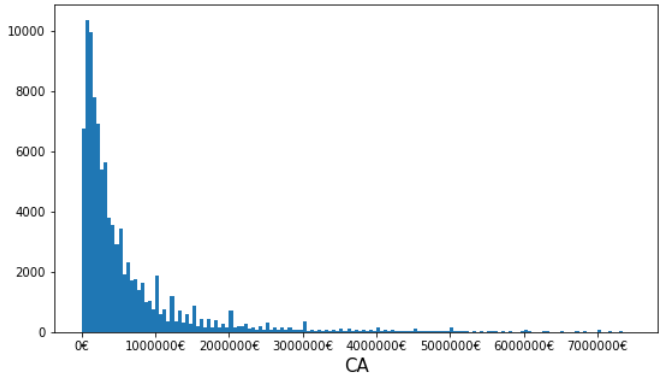

FIGURE 2.12 – Histogramme selon le Chiffre d'affaires

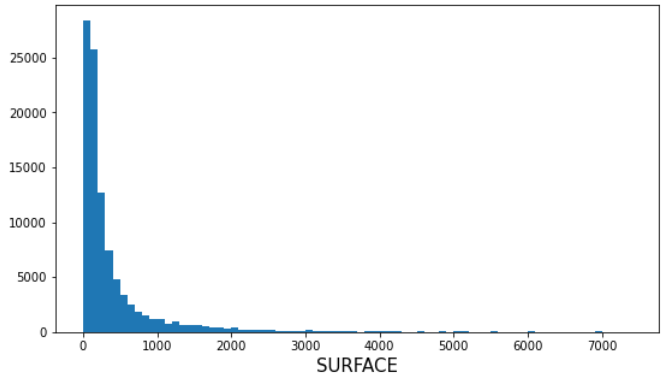

FIGURE 2.13 – Histogramme selon la surface assurée

L'assurance MRC couvre en grande majorité des petits commerces, on retrouve cela sur les deux histogrammes, mais il y a quand même présence en petite quantité de risques plus importants. La grande majorité a un chiffre d'affaires inférieur à un million d'euros et une surface inférieure à $750m^2$. Cela se retrouve évidemment dans la prime annuelle payée par l'entreprise l'année du sinistre qui est évidemment corrélée avec la taille de l'entreprise :

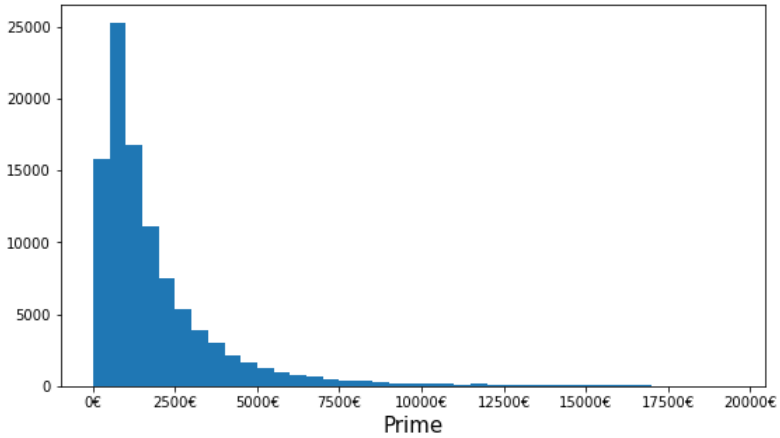

FIGURE 2.14 – Histogramme selon la prime

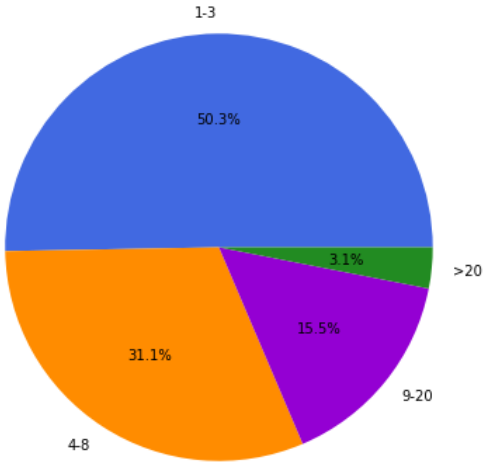

FIGURE 2.15 – Répartition des sinistres selon l'effectif du commerce

La grande majorité du portefeuille étudié est composée de petits commerces, d'artisans ... Il y a ainsi un petit effectif en général.

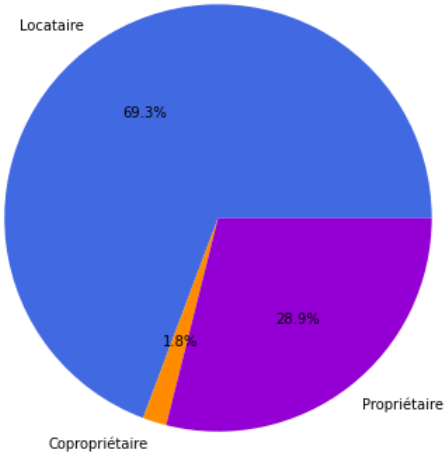

FIGURE 2.16 – Répartition des sinistres selon le type d'occupant

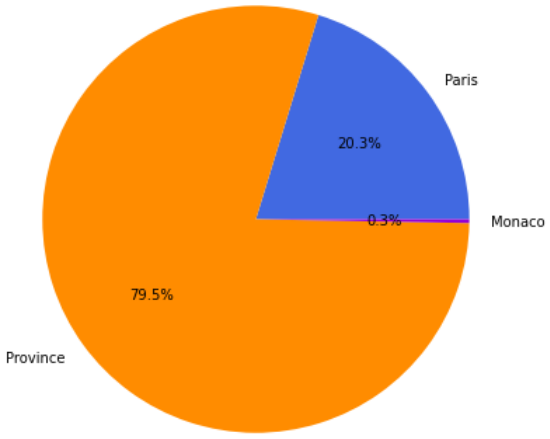

FIGURE 2.17 – Répartition des sinistres selon la localisation

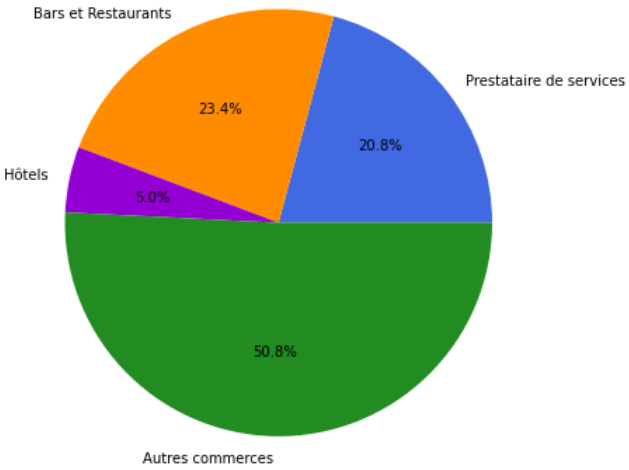

FIGURE 2.18 – Répartition des sinistres selon le groupement d'activité

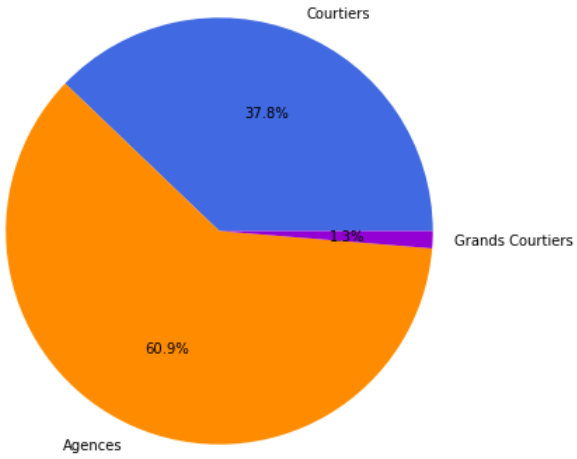

FIGURE 2.19 – Répartition des sinistres selon le réseau de distribution

Les occupants sont pour la grande majorité des locataires, ce qui est attendu pour un commerçant. Il y a très peu de copropriétaires. La presque totalité des sinistres est située en France et surtout en Province, ce qui est représentatif de notre portefeuille MRC en général. Le groupement d'activité est large et peu précis (seulement quatre catégories) mais il permet de se donner une idée du type de client. Comme on l'a vu, on est face à des petits risques en général, mais on peut déjà identifier les hôtels comme étant a priori de plus gros risques. Au niveau du réseau de distribution, les agents sont la principale source, mais il y a aussi des petits courtiers. Les grands courtiers sont moins représentés.

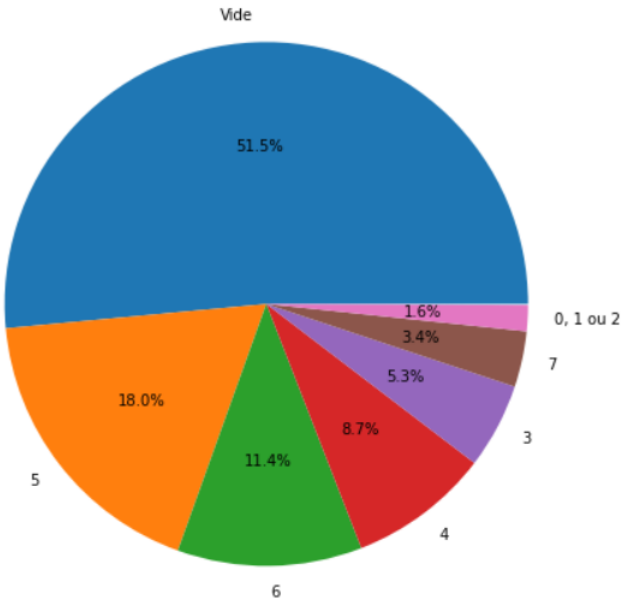

FIGURE 2.20 – Répartition des notes en année N

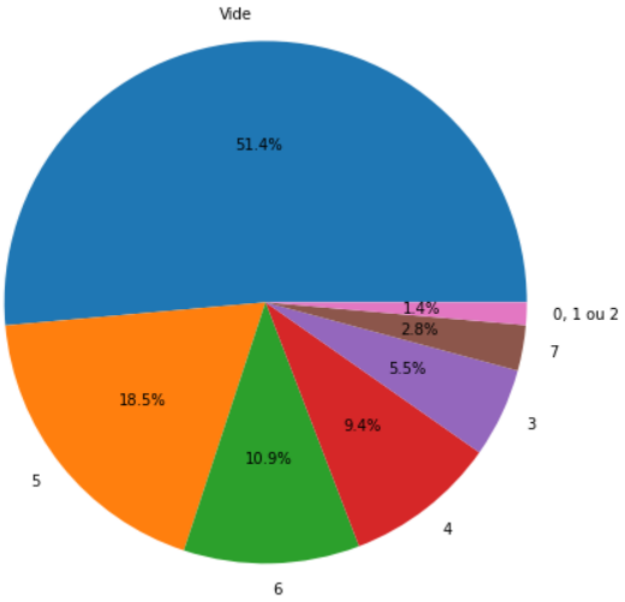

FIGURE 2.21 – Répartition des notes en année N-1

Pour les notes financières, on constate d'abord une certaine homogénéité entre les deux années. On a une majorité d'entreprises pour lesquelles on n'a pas accès à cette note financière. Pour les entreprises qui sont notées, la quasi-totalité est en bonne santé d'un point de vue financier. En effet, seulement 7% des entreprises ont une note inférieure ou égale à 3.

2.6 Données manquantes

Comme souvent, il y a présence de données manquantes dans la base de données. Étant donné que la majorité des algorithmes ne gèrent pas la présence de données manquantes, il est impératif de trouver une solution à ce problème.

La première idée serait de simplement d'enlever les lignes où il y a des données manquantes. Mais cette approche pose de nombreux problèmes. Elle amène d'abord à une réduction conséquente du nombre de lignes. Hors le nombre de lignes est un prérequis pour obtenir un bon modèle en général. L'incertitude baisse avec le nombre de lignes. De plus, cela peut amener un biais aux données dans le cas où la présence d'une donnée manquante ne soit pas totalement aléatoire. Pour illustrer cela, considérons le cas du chiffre

d'affaires. S'il manque le chiffre d'affaires alors cela veut sans doute dire que l'entreprise est très jeune. Donc le chiffre d'affaires est a priori faible. Ainsi, ignorer les données avec un chiffre d'affaires manquant aura tendance à sous-estimer la part des entreprises jeunes avec un chiffre d'affaires faible.

Il est donc clair qu'on ne va pas supprimer ces données. Une meilleure idée est ainsi de créer une nouvelle modalité (dans le cas d'une variable catégorielle) associée aux données manquantes. Une approche alternative souvent utilisée est de remplacer la donnée manquante par la moyenne (ou la médiane ou le minimum ou le maximum) dans le cas d'une donnée quantitative ou par la modalité majoritaire dans le cas d'une donnée catégorielle. On peut aussi avoir une approche métier comme dans l'exemple du chiffre d'affaires : on sait que la donnée est a priori faible et on peut ainsi remplacer par une valeur arbitraire. Mais dans le cas où la variable manquante est corrélée avec une ou plusieurs autres variables, on peut essayer de remplacer la donnée manquante par une estimation obtenue en appliquant un modèle de régression/classification à partir du reste de la base de données.

L'approche retenue est une méthode d'implantation multiple séquentielle[10]. Pour résumer cette méthode, prenons le cas où on a trois colonnes X, Y et Z. Imaginons qu'on a des données manquantes sur les variables X et Y. Il n'y a donc pas de données manquantes sur Z. Cette méthode se réalise en plusieurs cycles. Lors du premier cycle, on va commencer par remplacer les données manquantes de X en les estimant à partir de Y et Z en utilisant les lignes où il n'y a pas de données manquantes (sur X et Y) comme base d'entrainement. On obtient donc une nouvelle colonne X sans données manquantes. On va cette fois mettre à jour Y via une estimation à partir de X et Z. Le premier cycle est terminé : on n'a plus de données manquantes. On va maintenant estimer une nouvelle fois les données manquantes de X à partir de Y et Z, mais cette fois Y n'a plus de données manquantes. Une fois X mis à jour, on va mettre à jour Y de la même façon. On peut ainsi itérer un nombre de fois fixe ou jusqu'à ce que les estimations se stabilisent. Lors des différents cycles, l'ordre de mis à jour n'est pas fixe, on peut choisir de manière aléatoire la variable que l'on met à jour.

Comme algorithme de classification/régression, on a utilisé (selon la qualité de prédiction) soit un arbre CART soit un algorithme des k plus proches voisins. Pour que cette méthode soit efficace, il faut que la proportion de données manquantes pour une colonne ne soit pas trop élevée et il faut bien sûr que les variables soient au moins un peu corrélées entre elles.

Les méthodes CART sont plus détaillées dans la suite. On fait ici une brève présentation de l'algorithme des k-plus proches voisins : Soit un jeu de données constitué de d variables et de n individus $\{X_i \in \mathbb{R}^d\}_{1 \leq i \leq n}$.
Les k plus proches voisins d'un nouvel individu $x \in \mathbb{R}^d$ sont les k individus parmi les n individus du jeu de données qui minimisent la quantité $\|x - X_i\|$ où $\|\cdot\|$ est une norme définie sur $\mathbb{R}^d$. Dans le cas d'une variable continue, on prend généralement la norme euclidienne et dans le cas d'une variable discrète on prend la distance d'Hamming (annexe A).

On a donc réussi à remplir les données manquantes pour certaines variables. Avant de remplacer, on a bien sûr effectué des mesures sur une base de test où on connaissait la valeur. Voici un tableau faisant le bilan de cette méthode d'imputation :

Variable	Taux de vide	Erreur relative moyenne	AUC
Chiffre d'affaires	1.97%	18.4%	
Effectif	2.76%	35.6%	
Groupe d'activité	1.03%		0.87

TABLE 2.5 – Bilan : remplissage des données manquantes

L'erreur relative concerne le chiffre d'affaires et l'effectif, qui sont des variables quantitatives. L'erreur relative de 35.6% sur l'effectif peut paraître assez énorme, mais l'analyse du portefeuille montre que la quasi-totalité des entreprises a moins de 10 employés et une grande partie n'a pas plus de deux ou trois employés. Une erreur de 1 sur une entreprise ayant un effectif de 2 correspond à une erreur de 50 %. Le plus important, c'est qu'on arrive à discerner les petites, les moyennes et les grandes entreprises.
L'AUC concerne le groupe d'activité qui est une variable qualitative. Cette mesure est définie dans la partie 4.2.2. Dans les trois cas, les modélisations obtenues sont plutôt bonnes, et il est préférable de prendre les valeurs obtenues que de garder les valeurs vides.

Il reste encore des variables avec des données manquantes comme par exemple la date de création de l'entreprise ou la distance au pompier. Pour ces variables, il n'est pas logique d'essayer de prédire des valeurs à partir du reste de la base. Pour quand même pouvoir utiliser ces variables, on les a transformées en variables qualitatives en réalisant des regroupements. Par exemple, pour la date de création de l'entreprise, on a regroupé en plusieurs intervalles : [1900 ;1965],[1966 ;1980],[1981 ;1995],[1996 ;2005],[2006 ;2010] et [2011 ;2020].

2.7 Regroupement de modalités

Il y a beaucoup de variables catégorielles dans la base de données. Certaines ont beaucoup de modalités. Dans la plupart des algorithmes, les variables catégorielles sont transformées en variables binaires. Par exemple, imaginons qu'on a une variable catégorielle X qui peut prendre comme valeur A, B ou C. On va créer trois nouvelles variables : X_A, X_B et X_C qui vaudront 0 ou 1 selon la valeur de la modalité initiale. On voit bien qu'un grand nombre de modalités va faire exploser le nombre de colonnes de notre base de données. Avoir un trop grand nombre de colonnes peut être un frein à la qualité de notre modélisation.

Il est généralement conseillé[5] que pour éviter des problèmes de sur apprentissage, il faut avoir $n > k^2$ avec n le nombre de lignes (de l'ordre de 100 000 dans notre cas) et k le nombre de variables explicatives. Ce qui donne un k_{max} de l'ordre de 316. Ce problème peut être géré par des méthodes de pénalisation en fonction de l'algorithme utilisé. Mais il est peut-être utile de regrouper certaines modalités avant de commencer la modélisation. Surtout celles qui sont en faible proportion.

Dans notre base, cela concerne la forme juridique et le code NAF. La forme juridique est très précise, on sait par exemple s'il s'agit d'une SCP d'avocats ou bien un établissement public local culturel. Il y a plus de 100 modalités différentes et certaines modalités ne sont même pas représentées par plus de 10 individus. Le code NAF, lui est un code déterminé par l'INSEE selon le secteur d'activité. Encore ici, il y a plus de 100 modalités avec certaines très peu représentées.

Les regroupements de modalités se font généralement à l'aide de méthodes de clustering. On a, à la fois utilisé une méthode des k-means (Annexe C) et une méthode plus "pragmatique". En effet, on peut regrouper à la main certaines modalités, car elles sont proches d'un point de vue théorique. Par exemple, on peut se dire qu'une SCP de dentistes, qu'une SCP de médecins ou qu'une SCP de vétérinaires peuvent être regroupées ensemble. Ainsi, pour la forme juridique on arrive à 13 modalités et pour le code NAF, on obtient 45 modalités.

Détermination d'un seuil de gravité

Avant d'étudier la gravité des sinistres, il est nécessaire de définir ce qu'on entend par sinistre grave. C'est-à-dire définir un seuil à partir duquel un sinistre est considéré comme étant grave. Actuellement, ce seuil est fixé à 150 000 €. Ce seuil étant plutôt arbitraire, il va être question dans cette partie d'étudier si cette valeur est acceptable d'un point de vue théorique et si d'autres valeurs sont elles aussi possibles. En sachant que le seuil ne doit pas être pris trop élevé sous peine d'avoir trop peu de données pour la modélisation. On va considérer deux approches : la première se basant sur la charge du sinistre et la deuxième sur le taux de destruction défini ainsi :

$$TD = \frac{Charge\ du\ sinistre}{Engagement}$$

On commence par faire un bref rappel des principaux résultats[9] issus de la Théorie des Valeurs Extrêmes afin de mieux appréhender les différentes méthodes utilisées ensuite.

3.1 Théorie des valeurs extrêmes

3.1.1 Loi et convergence du maximum

Soit $(X_1, ..., X_n)$ une suite de variables aléatoires indépendantes et identiquement distribuées et soit F la fonction de répartition. Tandis que $\overline{F} = 1-F$ sera la fonction de survie. On définit le maximum $M_n = \max\{X_1, ..., X_n\}$ dont la fonction de répartition est donnée par :

$$\begin{aligned} F_n(x) &= P[X_1 \leq x; ...; X_n \leq x] \\ &= P[X_1 \leq x]...P[X_n \leq x] \\ &= [F(x)]^n \end{aligned}$$

La connaissance de la loi de X donne la loi de son maximum, mais en pratique on ne connaît pas la loi de X précisément. L'objectif est ainsi de savoir approcher la loi de M_n lorsque n tend vers $+\infty$ lorsqu'on ne connaît pas la loi de X.

On peut définir le point extrémal :

$$x^F = \sup\{x \ ; \ F(x) < 1\}$$

On introduit la notation suivante : $(X_{(1)}, ..., X_{(n)})$ telle que :

$$X_{(n)} \leq X_{(n-1)} \leq ... \leq X_{(1)}$$

Théorème 1 (Convergence)
Il existe une suite $(u_n)_n$ telle que $P[M_n \leq u_n]$ converge si et seulement si

$$\lim_{x \to x^F} \frac{\overline{F}(x)}{\overline{F}(x^-)} = 1$$

3.1.2 Domaine d'attraction du maximum

On dit que deux distributions F et G sont de même type si :

$$\forall x, F(ax + b) = G(x)$$

Par exemple, $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$ sont de même type. Les distributions sont identiques à un facteur d'échelle et un facteur de position près.

Théorème 2 (Fisher - Tippett)
S'il existe une suite $(a_n)_n$ de réels strictement positifs et une suite $(b_n)_n$ de réels telles que :

$$\lim_{n\to\infty} P\left[\frac{M_n - b_n}{a_n} \leq x\right] = G(x)$$

où G est une distribution non dégénérée. On dit que F appartient au domaine d'attraction de de G et on note $F \in D(G)$.

Alors G est une distribution des extrêmes généralisés (GEV).

Le théorème de Fisher-Tippett est l'équivalent du théorème central limite (Annexe B) dans le cas extrême (le TCL lui s'intéresse au comportement moyen). Si on arrive à trouver une normalisation linéaire alors on peut approximer la distribution par une distribution asymptotique sous condition d'avoir un échantillon assez grand. Contrairement au TCL où la distribution est unique (loi Normale), il y a plusieurs distributions possibles dans le cas extrême. De plus, contrairement au TCL, les variables de normalisation ne sont pas forcement connues. En pratique, il est donc plus difficile d'utiliser ce théorème.

Théorème 3 (Domaine d'attraction)
Soient F et G deux distributions telles que $x^F = x^G$.
Si $\lim_{x\to x^F} \frac{\overline{F}(x)}{\overline{G}(x)} = c \in \mathbb{R}_+^*$
Alors F et G sont dîtes équivalentes en termes de queue de distribution et elles ont le même domaine d'attraction.

Ainsi l'idée est de trouver le domaine d'attraction de notre loi F inconnue parmi les distributions de type GEV afin de modéliser le comportement asymptotique de notre distribution. Les distributions des extrêmes généralisés sont traitées plus spécifiquement dans la partie suivante.

3.1.3 Distribution des extrêmes généralisés (GEV)

Les distributions des extrêmes généralisés $GEV(\mu, \sigma, \xi)$ sont définies de cette manière :

$$G(x) = \begin{cases} \exp\left(-\left[1+\xi\left(\dfrac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right) & \text{si } \xi \neq 0 \\ \exp\left(-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right) & \text{si } \xi = 0 \end{cases}$$

où μ est le paramètre de position, σ le paramètre d'échelle et ξ le paramètre de forme ou l'indice de queue. ξ caractérise l'épaisseur de la queue de distribution : la queue étant épaisse quand $\xi > 0$, intermédiaire quand $\xi = 0$ et fine quand $\xi < 0$.

Les distributions les plus répandues sont :

Gumbel ($\xi = 0$) : $x \in \mathbb{R}$

$$\Lambda(x) = \exp(-\exp(-x))$$

Fréchet ($\xi > 0$) : $x > 0$, $\alpha > 0$,

$$\Phi_\alpha(x) = \exp(-x^{-\alpha})$$

Weibull ($\xi < 0$) : $x < 0$, $\alpha > 0$,

$$\Psi_\alpha(x) = \exp(-(-x)^{\alpha})$$

L'ensemble des lois limites du Théorème de Fisher-Tippett s'obtiennent par une transformation linéaire de l'une de ces trois distributions. Les distributions de Cauchy et de Pareto appartiennent ainsi au domaine de Fréchet : ce sont des distributions à queue lourde (à décroissance polynomiale).
Les distributions Normale, Exponentielle, Gamma et Log-Normale appartiennent au domaine de Gumbel : ceux sont des distributions à queue intermédiaire (à décroissance exponentielle).
La distribution Beta appartient elle au domaine de Weibull : c'est une distribution à queue fine.

De manière générale, si

$$h'(x) \xrightarrow[x \to x^F]{} \xi \qquad \text{où } h(x) = \frac{\overline{F}(x)}{f(x)}$$

Alors F appartient au domaine d'attraction d'une GEV de paramètre de queue ξ.

3.1.4 Excès au-delà d'un seuil

Théorème 4 (Excès-au delà d'un seuil)
Si F appartient au domaine d'attraction d'une GEV alors la loi des excès au-delà d'un seuil $(X - u|X > u)$ est une distribution de Pareto généralisée $GPD(\sigma, \xi)$. C'est-à-dire :

$$P\left[X - u > x | X > u\right] \xrightarrow[u \to x^F]{} \begin{cases} \left(1 + \frac{\xi x}{\sigma}\right)_+^{-1/\xi} & \text{si } \xi \neq 0 \\ \exp\left(\frac{-x}{\sigma}\right) & \text{si } \xi = 0 \end{cases}$$

De plus, si $0 < \xi < 1$ alors la Mean Excess Function $MEF(u) = E[X - u|X > u]$ est linéaire en u avec pour pente $\frac{\xi}{1-\xi}$.
Si $\xi \leq 0$ alors MEF aura une tendance "logarithmique" ou sera constante/décroissante.
Si $1 \leq \xi$ alors MEF n'est pas définie, car l'espérance est infinie.

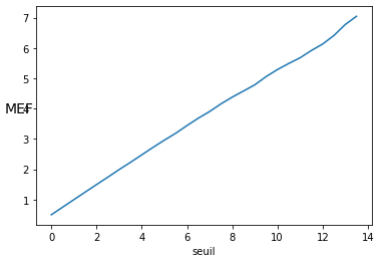

FIGURE 3.22 – MEF d'une Pareto de paramètre 3

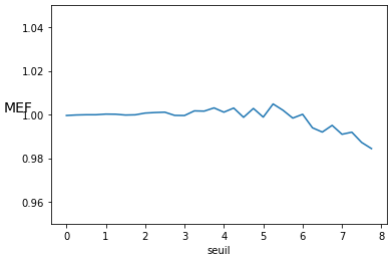

FIGURE 3.23 – MEF d'une Exponentielle

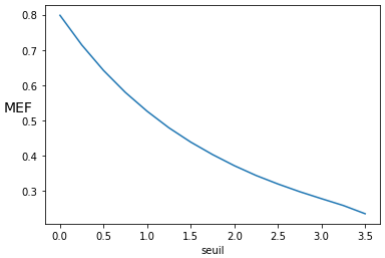

FIGURE 3.24 – MEF d'une loi Normale

En pratique, pour un seuil assez élevé (en général $x^F = +\infty$), les distributions GPD fournissent une bonne approximation de la queue de distribution. Les GEV caractérisent le maximum tandis que les GPD caractérisent la queue de distribution : les GPD apportent donc plus d'informations.

Théorème 5 (Stabilité des GPD)
Les lois GPD sont stables par troncature à gauche. C'est-à-dire :
Si $X \sim GPD(\sigma, \xi)$ alors $X - u|X > u$ suit encore une GPD de paramètre de queue ξ identique.

3.2 Détermination de seuil

3.2.1 Comportement de la queue et quantiles

La première approche à la détermination d'un seuil de gravité passe par l'analyse des quantiles. Intuitivement, les sinistres graves doivent représenter une faible part des sinistres, mais une grande part de la charge. C'est une approche très empirique.

On commence par regarder la queue de distribution :

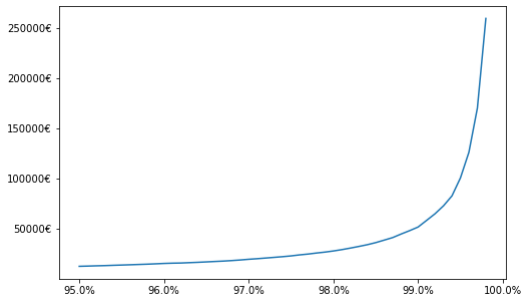

FIGURE 3.25 – Quantiles de la queue de distribution en Charge

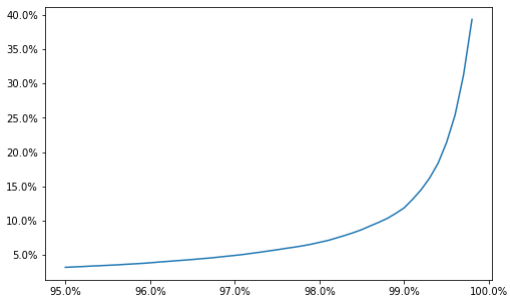

FIGURE 3.26 – Quantiles de la queue de distribution en TD

On voit que la charge et le taux de destruction commencent à augmenter de manière significative au alentours du quantile à 99%. Pour être plus précis, on peut s'intéresser aux variations inter-quantiles relatives. On va tracer :

$$\frac{x_i - x_{i-0.25\%}}{x_{i-0.25\%}} \text{ en fonction de } i$$

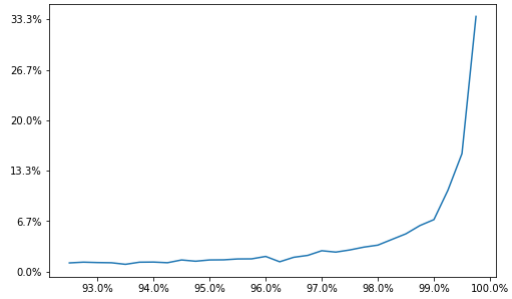

FIGURE 3.27 – Variation relative inter-quantiles (vision charge)

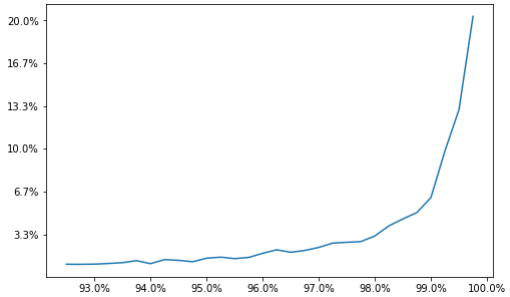

FIGURE 3.28 – Variation relative inter-quantiles (vision TD)

De manière équivalente, on commence à observer le comportement extrême de la queue à partir du quantile à 99%.

	Charge	TD
98.5%	36 441 €	8.7 %
99%	51 940 €	11.9 %
99.5%	100 794 €	21.4 %

TABLE 3.6 – Quelques valeurs de quantiles

Cette première approche suggérait un seuil entre 50 000€ et 100 000€ ou entre 12 % et 21 %.

3.2.2 Graphique quantiles-quantiles (Q-Q plot)

Les autres approches utilisent la théorie des valeurs extrêmes. On suppose donc que notre distribution appartient au domaine d'attraction d'une distribution des extrêmes généralisés de paramètre de queue ξ. La première étape passe par la détermination du type de queue ($\xi > 0$, $\xi < 0$ ou $\xi = 0$). Le q-q plot est le graphique obtenu en traçant ce nuage de point :

$$\left(X_{(i)}, F^{-1}\left(1 - \frac{i}{n}\right)\right)_{1 \leq i \leq n}$$

En remarquant que :

$$X \overset{\mathcal{L}}{=} F^{-1}(U) \text{ où } U \sim \mathcal{U}(0,1)$$

Les points obtenus doivent être alignés (même si les points sont obtenus par une transformation linéaire). Le q-q plot permet donc de vérifier si un échantillon suit une distribution de référence. L'objectif est de comparer notre échantillon à une distribution à queue intermédiaire ($\xi = 0$) comme la loi Exponentielle de paramètre $\lambda = \frac{1}{E[X]}$. C'est-à-dire ce nuage de point :

$$\left(X_{(i)}, -\frac{1}{\lambda} ln\left(\frac{i}{n}\right)\right)_{1 \leq i \leq n}$$

Si la courbe a une tendance concave alors notre distribution a une queue plus épaisse ($\xi > 0$). Si la courbe est plutôt convexe, la queue est plus fine ($\xi < 0$). Et si la courbe est linéaire alors la queue est équivalente ($\xi = 0$).

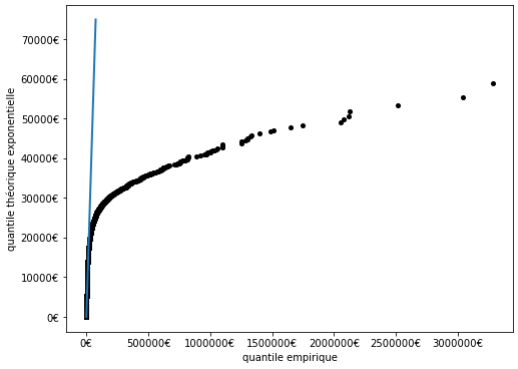

FIGURE 3.29 – q-q plot en charge

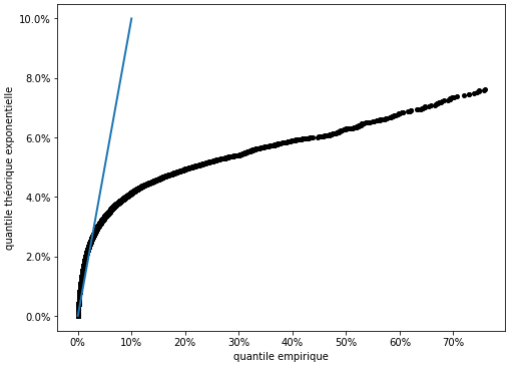

FIGURE 3.30 – q-q plot en TD

La droite bleue est la droite d'équation $y = x$. Les courbes (noires) obtenues sont clairement concaves : la queue est ainsi plus épaisse que la distribution exponentielle. **On est dans le domaine de Fréchet ($\xi > 0$).**

3.2.3 Mean Excess Function

Connaissant le signe de ξ (> 0), une première approche pour déterminer le seuil est la Mean Excess Function introduite dans le Théorème 4. En traçant la MEF empirique, on peut obtenir un seuil en prenant le plus petit seuil tel que la MEF semble linéaire à partir de ce seuil.

La MEF empirique est définie de cette manière :

$$\widehat{MEF}(X_{(k)}) = \sum_{i=1}^{k-1} \frac{X_{(i)} - X_{(k)}}{k-1}$$

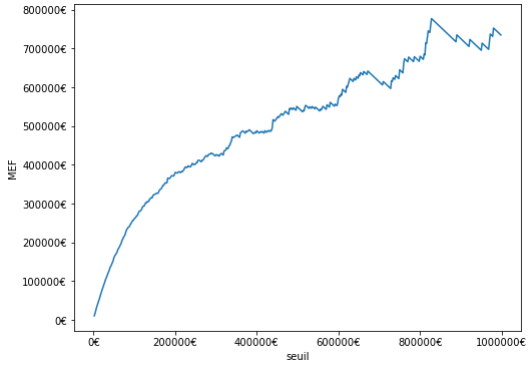

FIGURE 3.31 – MEF plot (charge)

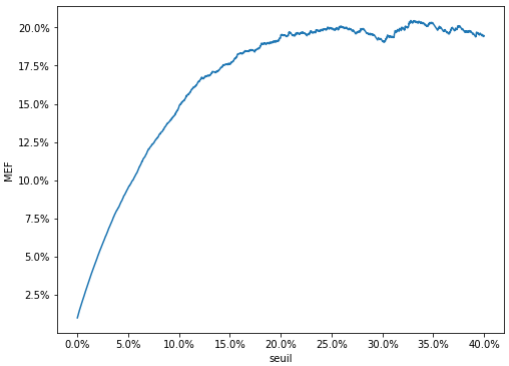

FIGURE 3.32 – MEF plot (TD)

Il est pertinent de ne pas regarder la zone la plus extrême du graphique, car cette zone n'est pas très représentative étant donné la faible densité de point dans cette zone (l'espérance n'étant pas robuste : l'estimation est peu précise). On voit très bien un comportement quasi linéaire sur le graphe de gauche à partir de la zone un peu avant les 200 000 €. La linéarité est moins claire sur le graphe de droite étant donné qu'à partir des 20%, la courbe semble constante. Encore une fois, cela peut être dû à la forte incertitude lors de l'estimation de l'espérance sur cette zone peu dense. Théoriquement, une MEF constante correspondrait à une distribution exponentielle. Or le q-q plot montre clairement qu'on n'est pas confronté à une telle distribution.

Pour optimiser le choix de ce seuil, pour chaque seuil, on peut réaliser une régression linéaire au-delà du seuil et regarder le R^2 et le σ de la régression associée. En effet, une régression de bonne qualité est associée à un R^2 proche de 1 et à un σ le plus faible possible. On peut donc décider de choisir un seuil en essayant de maximiser le R^2 et en en minimisant le σ de la régression linéaire effectuée. Pour cette régression, on retire évidemment les points les plus extrêmes qui sont peu robustes.

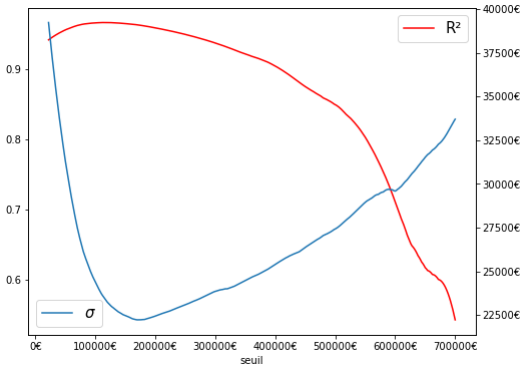

FIGURE 3.33 – R^2 et σ en fonction du seuil retenu (charge)

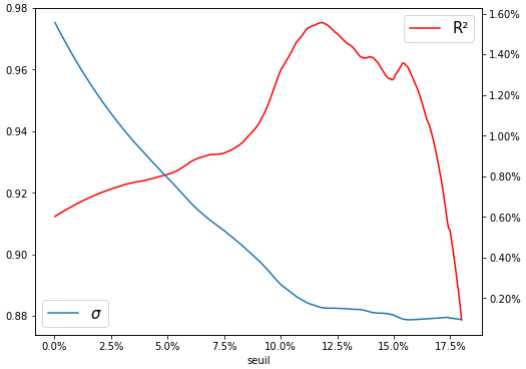

FIGURE 3.34 – R^2 et σ en fonction du seuil retenu (TD)

Sur le graphe de gauche :
— σ est minimal dans la zone 100 000€ - 300 000€
— R^2 est maximal dans la zone 0€ - 300 000€
Sur le graphe de droite :
— σ est minimal dans la zone 10% - 17.5%
— R^2 est maximal dans la zone 11% - 16%
En retenant des seuils à 150 000€ et 12% :

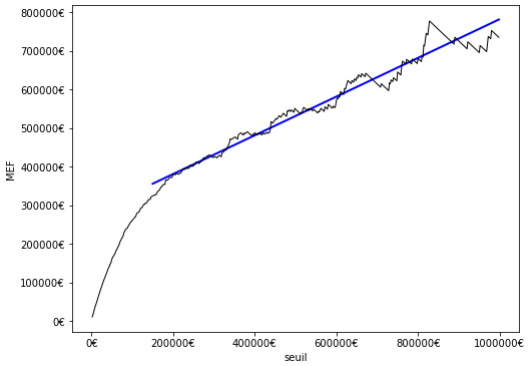

FIGURE 3.35 – MEF et régression (seuil à 150 000€)

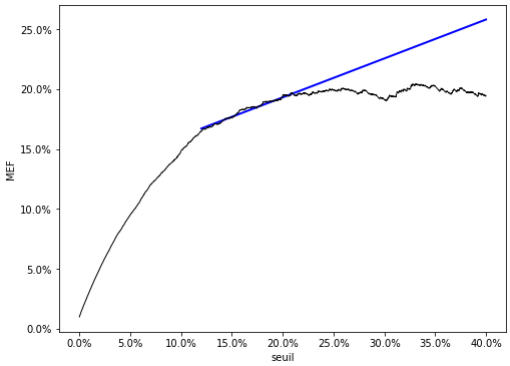

FIGURE 3.36 – MEF et régression (seuil à 12%)

Pour le seuil à 150 000€, la tendance est clairement linéaire et la régression est de bonne qualité. Pour le seuil à 12%, la droite de régression épouse bien la courbe au début, mais comme dit précédemment, la zone au-delà des 25 % est très mal modélisée due sans doute aux manques de points (seulement 0.4% des sinistres ont un taux de destruction supérieur à 25%).

3.2.4 Graphe des estimateurs de ξ

Une seconde approche de détermination de seuil consiste à utiliser la stabilité de ξ par troncature à gauche (Théorème 5). En effet, on a vu qu'une fois la zone extrême atteinte (seuil assez élevé), on peut modéliser notre queue de distribution par une GPD dont le paramètre de queueξ est une constante du seuil. En traçant le graphique des estimations $\widehat{\xi}$ pour différents seuils, on peut choisir comme seuil la plus petite valeur à partir de laquelle $\widehat{\xi}$ semble se stabiliser. En pratique, on préfère tracer de manière équivalente ξ en fonction du nombre d'excès k. Il existe de nombreux estimateurs de ξ : estimateur du maximum de vraisemblance, estimateurs paramétriques, estimateurs non paramétriques. On va se concentrer sur les deux estimateurs les plus utilisés : l'estimateur de Hill[6] et l'estimateur de Dekkers-Einmahl-de Hann[2] qui sont tous deux des estimateurs non paramétriques.

Estimateur de Hill ($\xi > 0$)

Cet estimateur est défini de cette manière :

$$\widehat{\xi_k} = \sum_{i=1}^{k} \frac{ln(X_{(i)}) - ln(X_{(k+1)})}{k}$$

Si $\frac{k}{n}$ est assez petit et k assez grand (et donc n très grand), on a convergence :

$$\sqrt{k}\left(\widehat{\xi_k} - \xi\right) \overset{\mathcal{L}}{\approx} \mathcal{N}(0, \xi^2)$$

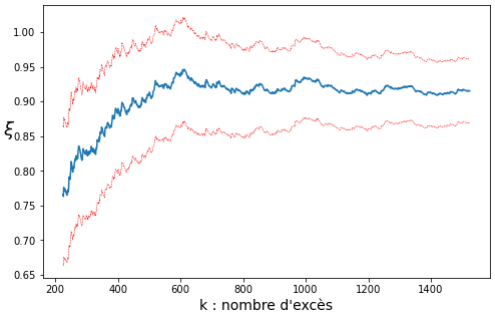

FIGURE 3.37 – Hill plot (charge)

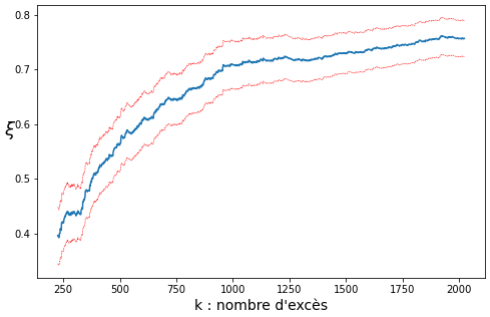

FIGURE 3.38 – Hill plot (TD)

Pour le graphe de gauche, tant qu'on prend un k supérieur à 600, la condition est satisfaite. Pour le graphe de droite, il semble y avoir une zone de stabilité entre 1000 et 1200.

Estimateur de Dekkers-Einmahl-de Hann ($\xi \in \mathbb{R}$)

Cet estimateur est défini de cette manière :

$$\widehat{\xi_k} = H_k(1) + 1 - \frac{0.5}{1 - \frac{H_k(1)^2}{H_k(2)}}$$

où :

$$H_k(n) = \sum_{i=1}^{k} \frac{\left(ln(X_{(i)}) - ln(X_{(k+1)})\right)^n}{k+1}$$

Si $\frac{k}{n}$ est assez petit et k assez grand (et donc n très grand), on a convergence :

$$\sqrt{k}\left(\widehat{\xi_k} - \xi\right) \overset{\mathcal{L}}{\approx} \mathcal{N}(0, 1 + \xi^2)$$

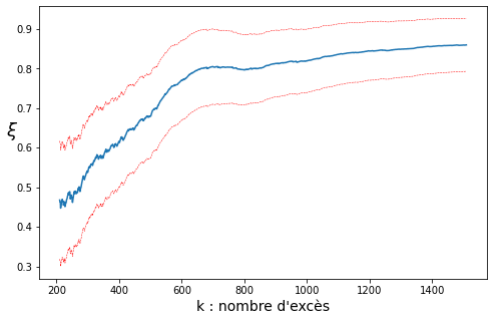

FIGURE 3.39 – DEdH plot (charge)

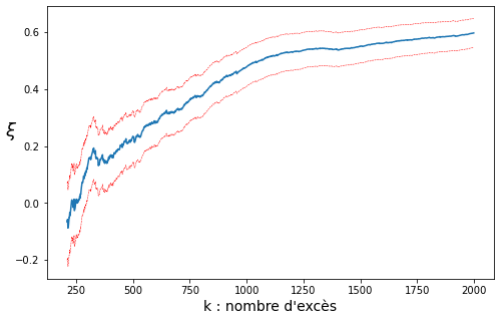

FIGURE 3.40 – DEdH plot (TD)

Pour le graphe de gauche, comme pour le Hill plot, tant qu'on prend $k > 600$, la condition de stabilité est satisfaite. Pour celui de droite, la zone entre 1000 et 1500 semble être un palier de stabilité. On peut remarquer que le $\widehat{\xi}_k$ est plus petit dans le cas du DEdH plot. En effet :

	Charge	TD
Hill plot	0.92	0.73
DEdH plot	0.81	0.55

TABLE 3.7 – Estimation de ξ dans la zone de stabilité identifiée

On a une incertitude plutôt grande (surtout dans le cas TD) sur l'estimation de ξ. Mais rappelons que l'on cherche juste un seuil et non pas une bonne estimation du paramètre de queue. On peut remarquer ici que l'hypothèse $0 < \xi < 1$ est bien satisfaite, pour l'instant on savait juste à l'aide du q-q plot que $0 < \xi$.

Utilisation de la MEF

On a vu que dans le cas où $0 < \xi < 1$, la MEF est linéaire avec pour pente : $p = \dfrac{\xi}{1-\xi}$. Un nouvel estimateur est donc :

$$\widehat{\xi} = \frac{\widehat{p}}{1+\widehat{p}}$$

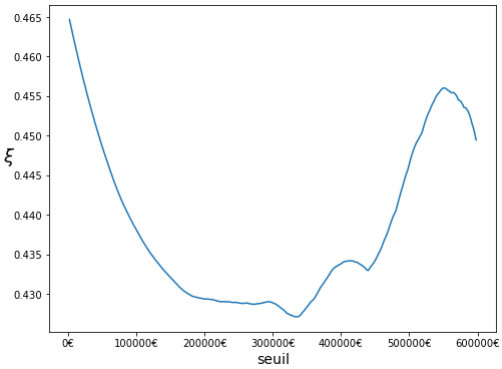

FIGURE 3.41 – Estimation de ξ (charge)

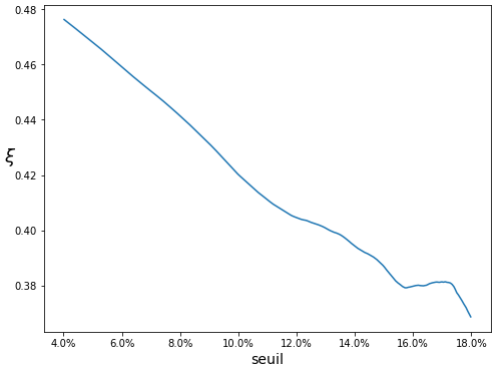

FIGURE 3.42 – Estimation de ξ (TD)

Les valeurs de ξ trouvées sont complètement différentes que celles trouvées précédemment et on n'observe pas vraiment de stabilité à part dans le graphe de gauche entre 150 000€ et 350 000€. Cette méthode n'étant pas la plus conseillée en pratique et l'incohérence des résultats font qu'on décide de ne pas utiliser cette méthode pour le choix définitif.

Bilan

	Charge	TD
nombre d'excès	800	1000
seuil associé	65 000€	11.8%

TABLE 3.8 – Seuils retenus avec les estimateurs de $\widehat{\xi}$

3.2.5 Gerstengarbe plot

On va utiliser[9] ici la méthode de Gerstengarbe[7], inspirée du test de tendance d'une série temporelle de Mann-Kendall. On construit les deux séries de différences :

$$\Delta_i = X_{(i+1)} - X_{(i)} \qquad \text{et} \qquad \widetilde{\Delta_i} = X_{(n-i)} - X_{(n-i+1)}$$

En notant :

$$n_k = \sum_{i=1}^{k-1} \mathbb{1}_{\Delta_i < \Delta_k} \qquad \text{et} \qquad \widetilde{n_k} = \sum_{i=1}^{k-1} \mathbb{1}_{\widetilde{\Delta_i} < \widetilde{\Delta_k}}$$

Ainsi que :

$$a_k = \frac{k(k-1)}{4} \qquad \text{et} \qquad b_k = \sqrt{\frac{k(k-1)(2k+5)}{72}}$$

On définit ces deux quantités :

$$H_k = \frac{\sum_{i=1}^{k} n_i - a_k}{b_k} \qquad \text{et} \qquad \widetilde{H_k} = \frac{\sum_{i=1}^{k} \widetilde{n_i} - a_k}{b_k}$$

L'idée est de trouver la valeur de k à partir de laquelle la série Δ n'est plus monotone. La méthode consiste à tracer $-H_k$ et $\widetilde{H_k}$ et de retenir le k associé au point d'intersection des deux courbes ainsi obtenues. Il est conseillé d'itérer ce procédé deux ou trois fois pour obtenir un seuil convenable.

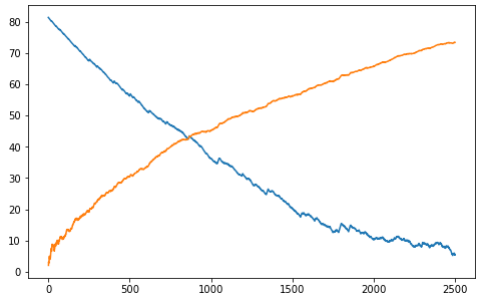

FIGURE 3.43 – Gerstengarbe plot (charge)

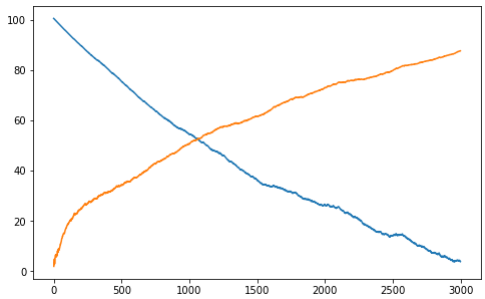

FIGURE 3.44 – Gerstengarbe plot (TD)

Après trois itérations, on obtient : à gauche $k = 860$ et à droite $k = 1070$. Ces deux seuils sont ainsi en accord avec les seuils obtenus précédemment.

3.2.6 Choix définitif

	Charge	TD
variations inter-quantiles	50 000-100 000€	12-21 %
MEF	150 000€	12%
stabilité de $\widehat{\xi}$	65 000€	11.8%
Gerstengarbe	60 000€	11.2%

TABLE 3.9 – Récapitulatif des différents résultats

Finalement un seuil en charge de 100 000€ (500 sinistres) et en Taux de destruction de 12% (987 sinistres) semblent être un bon compromis.

Prédiction des graves : modèle de propension

Dans cette partie, on commence par présenter brièvement ce qu'est un modèle de propension. On revient ensuite sur les points techniques importants à connaître au sujet de la méthode CART étant donné que cette méthode est la base des méthodes utilisées ensuite. On termine par présenter la modélisation du score de propension avec les méthodes Random Forest et Gradient Boosting.

4.1 Modèle de propension

Les modèles de propension sont généralement utilisés pour estimer l'influence d'une exposition sur une issue généralement binaire à l'aide d'observations. Dans notre cas, on a des informations représentant le risque d'un contrat : θ. Et l'on cherche à connaitre le score de propension défini de cette manière :

$$S(\theta) = \mathbb{P}[S > s | S > 0, \theta]$$

où S est la charge du sinistre et s le seuil de gravité du sinistre. On peut observer que dans le cas de l'approche par taux de destruction, s dépend de θ par l'intermédiaire de l'engagement : $s = 12\% \cdot \mathit{Engagement}$

Le score de propension est différent de la probabilité d'avoir un sinistre grave étant donné que la probabilité est conditionnée par le fait d'avoir un sinistre. À partir du score de propension, on obtient la probabilité d'avoir (au moins) un sinistre grave de cette façon :

$$\sum_{n=1}^{+\infty} \mathbb{P}[N = n|\theta] \cdot [1 - (1 - S(\theta))^n]$$

où on fait l'hypothèse que $S(\theta)$ et N sont indépendants conditionnellement à θ.

4.2 Outils mathématiques

4.2.1 Notations

Voici quelques notations utilisées dans la suite :
— k = nombre de variables explicatives
— n = nombre d'individus (ou de lignes)
— j = indice d'une variable (colonne)
— i = indice d'un individu (ligne)
— X_j = variable explicative d'indice j
— X_i = ensemble des variables explicatives pour l'individu i
— Y_i = valeur de la variable à expliquer pour l'individu i
— $f(X_i)$ désigne la prédiction du modèle pour l'individu i
— M = nombre d'arbres
— T désigne un arbre CART

4.2.2 Classification

Un problème de classification est un problème où l'on cherche à attribuer une catégorie Y_i à un individu X_i en se basant sur un ensemble de données sur d'autres individus dont on connait la catégorie. S'il y a p catégories, cela revient à cherche f à valeurs dans $[\![1, p]\!]$ telle que $f(X_i) = Y_i \in [\![1, p]\!]$. Si p=2, on parle de classification binaire en on prend comme catégorie 0 et 1 plutôt que 1 et 2.

On peut construire une matrice de confusion :

	$f(X_i) = 0$	$f(X_i) = 1$
$Y_i = 0$	vrais négatifs	faux positifs
$Y_i = 1$	faux négatifs	vrais positifs

En calculant les quatre effectifs de la matrice, la matrice de confusion permet d'indiquer la qualité de prédiction du modèle. On cherche à maximiser le nombre de vrais positifs et de vrais négatifs tandis qu'on souhaite minimiser le nombre de faux positifs et de faux négatifs.

On peut définir :

$$Sensibilite = \frac{VP}{FN + VP} \quad (= Rappel)$$

$$Specificite = \frac{VN}{VN + FP}$$

$$Precision = \frac{VP}{VP + FP}$$

Dans un problème de classification, on cherche à déterminer les positifs (les 1). La spécificité est le pourcentage de négatifs bien classés par le modèle. Même si on cherche à trouver les positifs, il est aussi important de ne pas trop se tromper sur les négatifs. La précision est le pourcentage de positifs prédits qui sont effectivement positifs. La précision informe sur la fiabilité du modèle concernant la modalité 1. La sensibilité (ou rappel) est le pourcentage de vrais positifs qui sont bien classés par le modèle. La sensibilité donne le pouvoir prédictif de notre modèle concernant la modalité 1.
Quelque soit le modèle, il faut trouver un compromis entre sensibilité, spécificité et précision. On s'intéresse souvent en particulier à la sensibilité et à la spécificité. En général, elles sont asymétriques : augmenter l'une des deux aura tendance à diminuer l'autre. Pour trouver un compromis, on peut utiliser la courbe **ROC** (receiver operating characteristic) qui est la courbe obtenue en traçant la sensibilité en fonction de $1 - specificite$ en faisant varier le critère de classification.

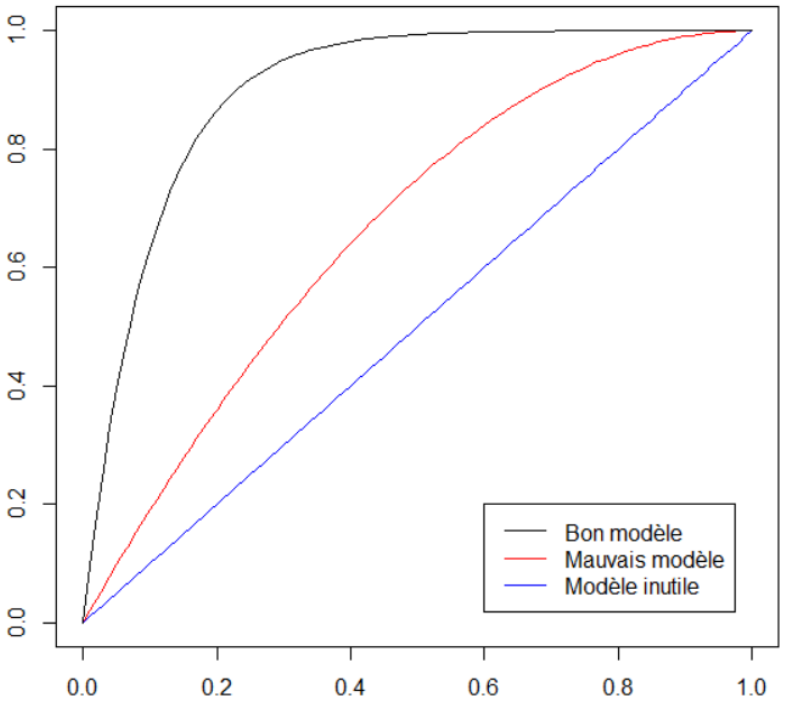

FIGURE 4.45 – Exemples de courbes ROC

Sur le graphique, on remarque que plus la courbe associée à notre modèle est "éloignée" de la courbe bleue alors plus le modèle est de bonne qualité. C'est ce qui amène à utiliser le critère **AUC** (area under curve), aussi appelé Statistique-C. C'est tout simplement l'aire en dessous de la courbe ROC. On cherche à avoir cette aire la plus proche de 1 possible. L'AUC permet de mesurer la qualité du modèle et surtout de comparer deux modèles entre eux.

Une autre mesure intéressante est le F_β-score qui se concentre sur la précision et le rappel :

$$F_\beta = \frac{(1+\beta^2) \cdot precision \cdot rappel}{\beta^2 \cdot precision + rappel}$$

Le score le plus utilisé est le F_1-score qui équipondère précision et rappel (moyenne harmonique). Le F_β**-score** mesure l'efficacité lorsqu'on attache β fois plus d'importance au rappel qu'à la précision. On choisit $\beta < 1$ si l'on souhaite pénaliser les faux positifs et $\beta > 1$ si l'on souhaite pénaliser les faux négatifs.

Le dernier indicateur utilisé pour mesure la performance d'un modèle de classification est le κ **de Cohen** :

$$\kappa = \frac{P_{accord} - P_{hasard}}{1 - P_{hasard}}$$

où $P_{accord} = \dfrac{VP + VN}{VP + VN + FN + FP}$ est la proportion d'accord entre le modèle et la réalité

et $P_{hasard} = \dfrac{(VP+FP)(VP+FN)}{VP+VN+FN+FP} \cdot \dfrac{(FN+VN)(FN+FP)}{VP+VN+FN+FP}$ est la probabilité que le modèle soit juste de manière aléatoire.

Le κ est un indicateur sur l'accord de notre modèle avec la réalité par rapport à un modèle complètement aléatoire. Si $\kappa = 0$, le modèle a la même performance qu'un modèle qui serait purement aléatoire. Si $\kappa < 0$, le modèle est moins performant que le hasard. Si $\kappa > 0$, le modèle est meilleur que le hasard. Plus κ est proche de 1 et plus le modèle sera considéré comme performant. On considère que le modèle commence à être utile quand $\kappa > 0.5$.

4.2.3 Surapprentissage

Lorsqu'on met en place un modèle d'apprentissage statistique, on cherche à le généraliser. C'est-à-dire à l'utiliser sur de nouvelles données à des fins pré-

dictives. La qualité du modèle se mesure ainsi à sa qualité de prédiction sur une base de données indépendante. Généralement, un modèle très complexe avec beaucoup de paramètres amène à un problème de surapprentissage. Le modèle surinterprète les données en se concentrant sur ce qui relève du détail plutôt que de déterminer la tendance globale. Il est donc d'usage[5] de séparer notre base de données en deux bases : la **base d'entrainement** sur laquelle on optimisera le modèle et la **base de test** sur laquelle on vérifiera la qualité de généralisation du modèle. Pour chaque méthode utilisée, il est ainsi nécessaire de combattre le nombre trop grand de paramètres du modèle. La plupart du temps, cela nécessite des hyperparamètres, c'est-à-dire des paramètres qui ne sont pas estimés, mais qui doivent être fixés à l'avance. Ces hyperparamètres vont forcer le modèle à ne pas être trop complexe. Ces hyperparamètres dépendent du type de modèle choisi. Comme ce choix est totalement arbitraire, il est nécessaire de mettre de côté une partie de notre base pour regarder quel hyperparamètre est le meilleur choix. La partie de la base qui est séparée est appelée **base de validation**. On réalise généralement cette opération de séparation plusieurs fois. Imaginons qu'on décide de séparer notre base en trois. On va estimer le modèle choisi trois fois : à chaque fois en retirant une des parties. C'est ce que l'on appelle la **validation croisée**.

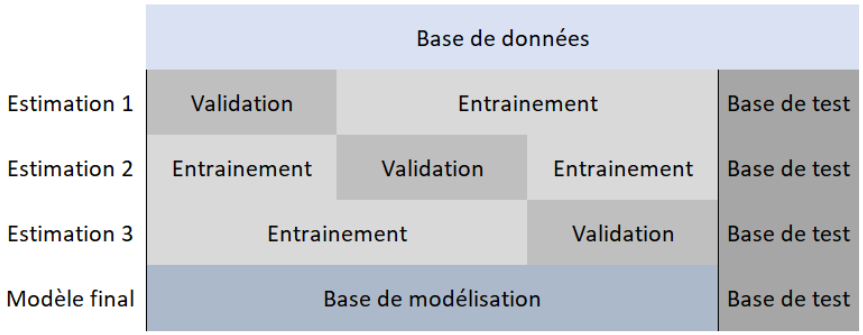

FIGURE 4.46 – Illustration de la séparation de la base et de la validation croisée

La validation croisée permet d'avoir un indicateur de la qualité des prédictions d'un modèle. Pour choisir la meilleure valeur pour l'hyperparamètre, il faut tester de la manière la plus exhaustive possible les valeurs possibles de l'hyperparamètre. En général, on a plusieurs hyperparamètres.
Par exemple, si on a deux hyperparamètres λ_1 et λ_2 que l'on cherche dans

Algorithm 1 Validation croisée

1: Séparation de la base en K échantillons de manière aléatoire $\Omega = \bigcup_{p=1}^{K} \Omega_p$
2: Pour p allant de 1 à K :
3: Entrainement du modèle sur $\Omega \backslash \Omega_p$
4: Calcul de l'erreur sur Ω_p
5: Calculer la moyenne des erreurs

deux ensembles discrets E_1 et E_2. On va réaliser une validation croisée pour chaque couple $(\lambda_1, \lambda_2) \in E_1 \times E_2$. On retiendra le couple qui donne le meilleur score. Cette méthode est appelée **grid search**.
Cette méthode peut vite devenir couteuse en temps de calcul. Par exemple, pour la recherche des hyperparamètres de l'un des modèles, on avait $card(E_1) = 2$, $card(E_2) = 10$, $card(E_3) = 15$, $card(E_4) = 3$ et $k = 5$ (validation croisée). Ce qui donne $2 \cdot 10 \cdot 15 \cdot 3 \cdot 5 = 4500$ modèles à estimer. En sachant que chaque modèle est lui même composé de plusieurs sous modèles (des arbres CART).

4.2.4 Données déséquilibrées

Un problème auquel on peut être confronté en classification est la prédominance d'une modalité sur l'autre. Dans notre cas, les sinistres graves sont sous-représentés par rapport aux sinistres non graves. Un grand déséquilibre sur la variable prédictive implique un biais sur notre modèle puisque lorsqu'on optimise notre modèle, on va plutôt avoir tendance à chercher à bien modéliser la variable prédominante, car cela donnera l'illusion d'une bonne performance. Il est donc nécessaire de prendre en compte ce déséquilibre.

Il existe plusieurs méthodes pour faire face au déséquilibre d'un jeu de données[3]. On peut les regrouper en deux groupes : il y a celles qui cherchent à "améliorer" le jeu de données et celles qui cherchent à "améliorer" la modélisation. Dans la première catégorie, on cherche à modifier le jeu de données initial. On peut soit créer de nouvelles lignes avec des méthodes de suréchantillonnage soit enlever des lignes avec des méthodes de sous-échantillonnage de façon à rendre le jeu de données d'entrainement plus équilibré. Le suréchantillonnage est à privilégier quand on n'a pas suffisamment de données et inversement, le sous-échantillonnage quand on a beaucoup de données. La deuxième catégorie de méthode ne modifie pas le jeu de données, on va prendre en compte le déséquilibre lors de la construction du modèle. Une première approche est le **cost-sensitive-learning** (approche sensible au coût). Elle passe par la modification de la fonction erreur (Gini ou Entropie dans

notre cas, voir 4.3). Le plus souvent, on introduit une pondération des valeurs : une erreur sur un certain individu pourra coûter plus qu'une erreur sur un autre. Le choix de cette pondération est arbitraire, c'est un hyperparamètre à optimiser par validation croisée (voir 4.2.3). L'introduction de cette pondération fait que l'on ne peut plus interpréter le score de propension comme une probabilité, car en quelque sorte, on augmente artificiellement la probabilité d'être grave en faisant cela. Mais cela est nécessaire pour la qualité de l'algorithme de prédiction. Pour obtenir une probabilité, il faudrait trouver une méthode de renormalisation du terme obtenu.

4.3 CART

La méthode CART (classification and regression trees) est une méthode permettant de construire un arbre binaire (annexe D) de décision[5]. Imaginons que l'on a k variables explicatives $X_1, ...; X_k$ et que l'on cherche à expliquer une variable Y. La méthode CART consiste en la construction d'une séquence de nœuds. Chaque nœud étant associé à une des k variables explicatives et à un seuil qui permet de séparer l'ensemble en deux sous-ensembles. Le premier noeud de la séquence, appelé racine reçoit en entrée l'ensemble de l'échantillon de départ. Les nœuds suivants ne reçoivent qu'un sous-ensemble de l'échantillon total.

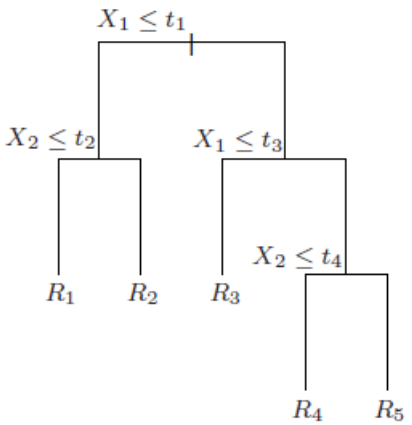

FIGURE 4.47 – Illustration d'un arbre de décision binaire

Pour définir un nouveau nœud dans la séquence, il est nécessaire de définir un critère de division qui permet à la fois de choisir quelle variable utiliser

(splitting variable) et quel seuil retenir (split point). Dans le cas d'une variable continue, ce critère repose sur des mesures usuelles comme l'erreur quadratique moyenne. Dans notre cas, pour une classification ce critère repose sur la notion d'hétérogénéité : on cherche à grouper les individus de manière homogène selon Y. Un critère (ou indice) d'hétérogénéité doit donc être minimal quand le nœud est homogène, c'est-à-dire que tous les individus du nœud ont la même valeur pour Y. Il doit être de plus en plus grand quand les valeurs de Y sont très dispersées. Le pire cas étant celui où il y a une proportion équirépartie dans chaque nœud.

Les deux critères les plus utilisés sont :

— L'indice de Gini :

$$H = 2p(1-p)$$

où p est la proportion de 1 dans le noeud considéré.
Si dans le noeud, il y a trois 0 et deux 1 alors l'indice de Gini vaudra

$$2 \cdot \frac{2}{5} \cdot \frac{3}{5}$$

— L'entropie croisée ou "log-loss" :

$$S = -p \cdot log_2(p) - (1-p) \cdot log_2(1-p)$$

En reprenant l'exemple précédent, l'entropie vaut :

$$-\frac{2}{5} \cdot log_2(\frac{2}{5}) - \frac{3}{5} \cdot log_2(\frac{3}{5})$$

Ces deux critères sont minimaux quand $p \in \{0, 1\}$ et maximaux quand $p = \frac{1}{2}$.

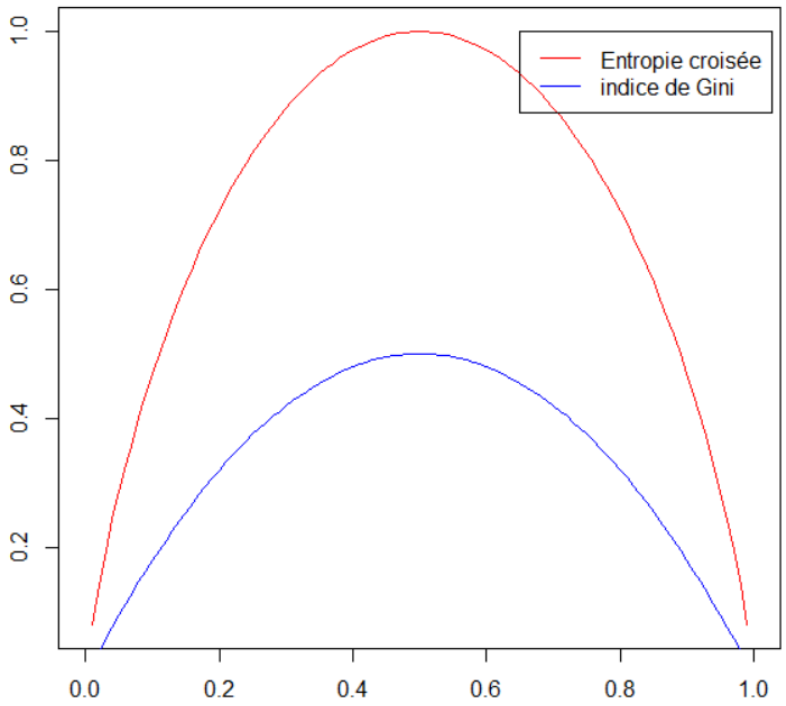

FIGURE 4.48 – Gini et entropie croisée pour un nœud en fonction de p

En plus de bien représenter la notion d'hétérogénéité, ces deux indices sont utilisés, car ils sont facilement optimisables en pratique de manière numérique.

Ainsi pour chaque variable explicative X_i, on peut déterminer le seuil qui maximise le gain d'hétérogénéité selon un des deux critères précédents. Une fois fait cela pour chaque X_i, on retient celle qui a gain d'hétérogénéité le plus élevé. Le nœud suivant est ainsi créé. Le nombre d'individus étant fini, il y a un nombre fini de partitions possibles. L'algorithme est convergent : il converge vers un arbre appelé arbre maximal. Le choix de l'un ou l'autre de ces deux indices étant arbitraires, ce choix se fera par validation croisée.

Le gain d'hétérogénéité est donné par :

$$\frac{N_{parent}}{n} \cdot \left(H_{parent} - \frac{N_{droit}}{N_{parent}} \cdot H_{droit} - \frac{N_{gauche}}{N_{parent}} \cdot H_{gauche} \right)$$

avec H l'indice de Gini ou l'entropie et N l'effectif associé au noeud.

L'utilisation d'une approche cost-sensitive-learning revient à changer les effectifs N en les pondérant par la pondération fixée.

Exemple :
On prend le cas où on a 7 individus, 2 variables explicatives et une variable à expliquer binaire :

X_1	X_2	Y
0.5	1	0
0.75	0	0
0	1	1
1	0.25	0
0.25	0.25	1
1	0.5	1
0.5	0.75	0

Pour cet exemple, on va utiliser l'indice de Gini et on va juste construire le premier split de l'arbre. On trace donc le gain en hétérogénéité selon le seuil pour chacune des deux variables :

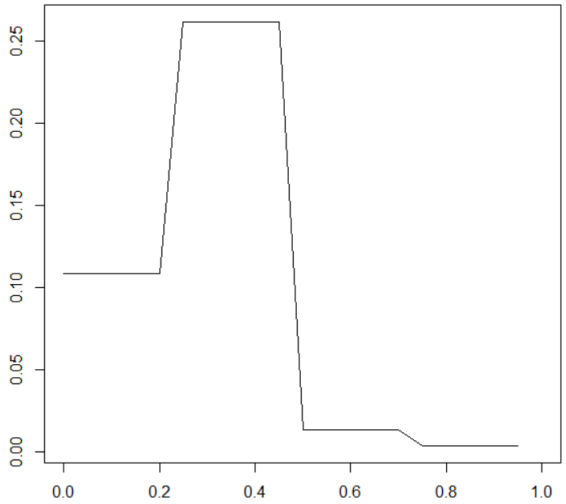

FIGURE 4.49 – Gain en hétérogénéité selon le seuil en X_1

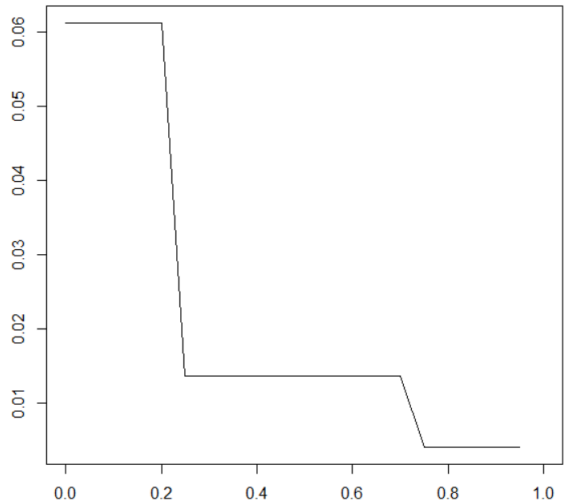

FIGURE 4.50 – Gain en hétérogénéité selon le seuil en X_2

Le gain en hétérogénéité est maximal sur $[0.25; 0.5[$ (graphe de gauche). Le split idéal est donc une séparation selon $X_1 \leq 0.375$ et $X_1 > 0.375$. Le choix de 0.375 est arbitraire, on pourrait décider de prendre n'importe quel nombre dans $[0.25; 0.5[$. On prend généralement la médiane.

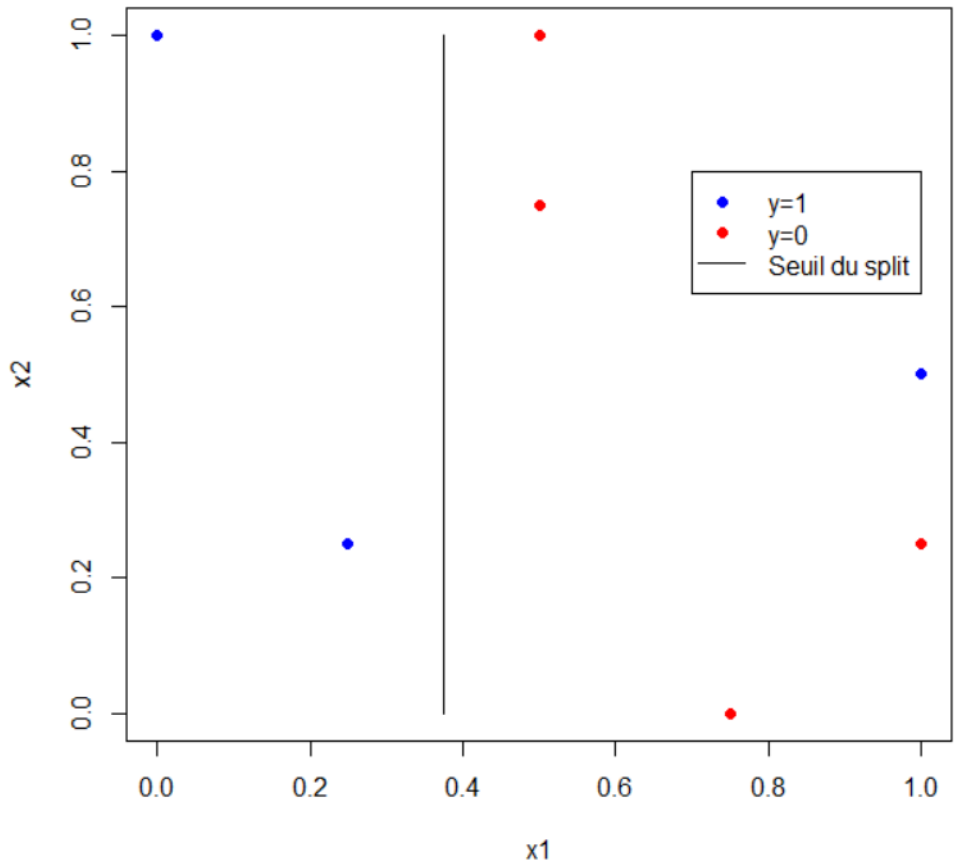

FIGURE 4.51 – Illustration du split effectué

On remarque bien qu'on a réussi à bien séparer les deux groupes : on passe d'un seul groupe avec trois 1 et quatre 0 à un groupe avec deux 1 et un autre avec quatre 0 et un seul 1. Le split permet d'isoler les deux points les plus à gauche : cette répartition est optimale au sens de l'indice de Gini.Si on s'arrête à ce stade, lors de la prédiction on affectera 1 à un nouvel individu si $X_1 \leq 0.375$ et 0 si $X_1 > 0.375$.

Pour éviter le surapprentissage associé à l'arbre maximal, on a plusieurs possibilités comme par exemple :

— Fixer une profondeur maximale à l'arbre
— Fixer un nombre de noeuds maximal
— Fixer un seuil minimal à la perte d'hétérogénéité
— réaliser du "pruning", c'est-à-dire couper l'arbre maximal

Le modèle Random Forest retenu dans la suite utilise une méthode de pruning. Pour réaliser du pruning, on agit comme cela :

— On construit l'arbre maximal de la manière décrite précédemment.
— Pour un arbre T, on définit $H_\alpha(T) = H(T)+\alpha\cdot|T|$ où $|T|$ est le nombre de nœuds terminaux de l'arbre T et où $H(T)$ est l'hétérogénéité des nœuds terminaux.
— Pour $\alpha \in \mathbb{R}$, on sélectionne le sous-arbre de l'arbre maximal qui minimise H_α

De cette manière on pénalise les nœuds qui sont peu efficaces en termes de gain d'hétérogénéité en prenant en compte leur taille.

Illustrons cela avec l'exemple précédent :

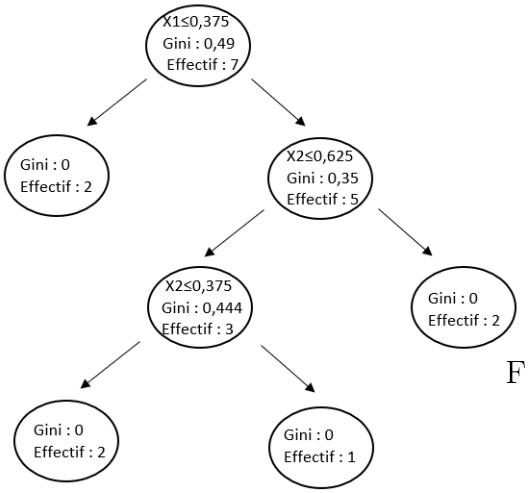

FIGURE 4.52 – arbre maximal

FIGURE 4.53 – Sous-arbre 1

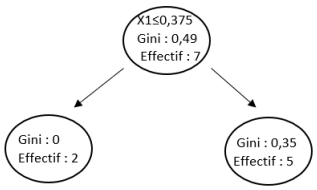

FIGURE 4.54 – Sous-arbre 2

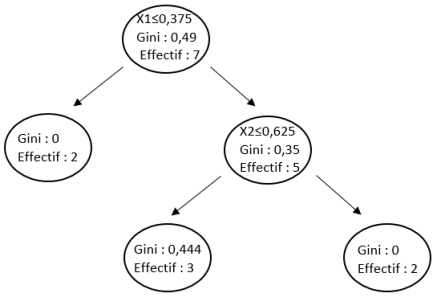

FIGURE 4.55 – Sous-arbre 3

On a :
— H_α(arbre max)= 4α
— H_α(Sous-arbre 1)= $0.49+\alpha$
— H_α(Sous-arbre 2)= $0.25+2\alpha$
— H_α(Sous-arbre 3)= $0.19+3\alpha$

Ainsi, si on prend α=0.2 alors l'arbre retenu est le sous-arbre 2 et si on décide de prendre α=0.3 alors on retient le sous-arbre 1. Augmenter α réduit bien la complexité de l'arbre.

Pour finir, on peut affecter une prédiction à partir de l'arbre de décision construit en l'appliquant à un nouvel individu et en affectant pour valeur de Y, la valeur la plus représentée dans le nœud obtenu. De manière générale, la proportion de 1 dans le nœud affecté sera la probabilité estimée que le nouvel individu soit de la catégorie 1. On peut aussi estimer une probabilité d'appartenance à chaque classe en utilisant les proportions dans le nœud terminal d'arrivée.

4.4 Random Forest

4.4.1 Présentation

Random Forest[1] ou Forêt aléatoire en français est une méthode dérivée de la méthode CART par bagging (bootstrap aggregating). La méthode de bagging consiste en la création de M modèles (hyperparamètre) obtenus de manière indépendante sur des échantillons bootstrap. Un échantillon bootstrap étant un ré-échantillonnage avec remise de la même taille que l'échantillon de départ.

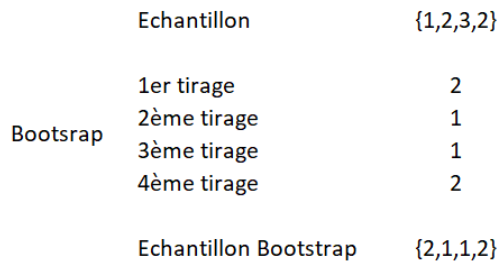

FIGURE 4.56 – Création d'un échantillon bootstrap

Dans le cas de Random Forest, le modèle de base est le modèle CART. Random Forest ajoute une subtilité supplémentaire : lors de la séparation d'un noeud en deux, on ne retient que m variables parmi les k variables disponibles de manière aléatoire. On peut prendre $m = \sqrt{k}$.

Algorithm 2 Random Forest

modèle={}
Pour p allant de 1 à M :
    E_p=Échantillon bootstrap
    Construction de l'arbre T_p sur E_p avec :
        pour chaque segmentation :
            choisir m variables de manière aléatoire
            création du meilleur noeud à partir des m variables
     On ajoute l'arbre T_p à modèle

La classification se fait en appliquant chaque arbre au nouvel individu et en procédant à un vote.

4.4.2 Application

On a donc cherché à estimer deux modèles Random Forest : le premier qui cherche à prédire les sinistres supérieurs à 100000 € et le second, les sinistres avec un taux de destruction supérieur à 12%.

Après l'application de Grid Search de la validation croisée pour trouver les hyperparamètres, le modèle pour les sinistres en charge est un Random Forest à 200 arbres, utilisant l'entropie, avec une profondeur maximale de 8 et un paramètre de pruning de 0.004. Le modèle pour les sinistres en taux de destruction est un Random Forest à 200 arbres, utilisant l'entropie, avec une profondeur maximale de 9 et un paramètre de pruning de 0.0007.

	Modèle en charge	Modèle en TD
Nombre d'arbres	200	200
Indice d'hétérogénéité	Entropie	Entropie
Profondeur maximale	9	8
Paramètre de pruning	0.004	0.0007
Pondération des graves	équilibrée	équilibrée

TABLE 4.10 – Récapitulatif des hyperparamètres

On constate que les deux modèles ont des hyperparamètres semblables à part pour le paramètre de Pruning. Dans les faits, l'indice hétérogénéité choisi n'influence que très peu sur la qualité du modèle. Pour le nombre d'arbres, tant qu'on prenait un nombre supérieur à 150, la qualité du modèle restait à peu près la même. La pondération équilibrée signifie ici que le coefficient est tel que les graves ont la même importance que les non graves. Par exemple, s'il y a 1 sinistre grave pour 100 sinistres alors les graves auront une pondération de 99 de manière à équilibrer l'importance des effectifs. Ce choix a été retenu, car les poids optimaux selon le grid search étaient proches du poids donné par cette méthode.

nombre d'arbres	100	150	200
profondeur maximale	7	5	7
α	0.005	0.003	0.001
Pondération	150	180	190
AUC (Gini)	68.5%	68.5%	69.1%
AUC (Entropie)	68.4%	68.7%	69.3%

TABLE 4.11 – Illustration de l'influence du choix de l'indice

Les exemples du tableau illustrent bien la similarité entre les deux indices, le choix de l'indice ne semble pas avoir un impact significatif sur le modèle. On a donc retenu l'entropie de manière arbitraire.

Le paramètre de pruning a permis de pouvoir augmenter la profondeur des arbres tout en évitant le surapprentissage. Par exemple, pour le modèle en charge on obtient une AUC sur la base test de 66% sans paramètre de pruning ($\alpha = 0$) contre une AUC de 71% pour le modèle retenu.

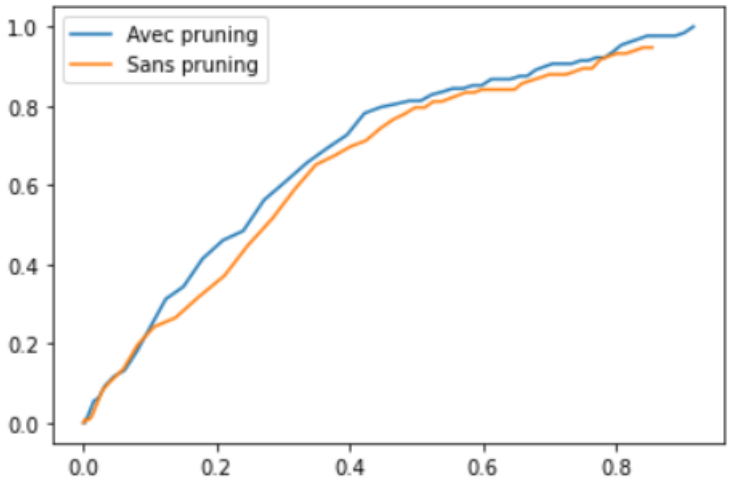

FIGURE 4.57 – Courbes ROC avec et sans pruning

Si on fixe le seuil de classification à 50% de manière classique, on obtient ces deux matrices de confusion (sur ma base de test) :

Charge	$f(X) = 0$	$f(X)) = 1$
$Y = 0$	14960	4723
$Y = 1$	65	63

TD	$f(X) = 0$	$f(X) = 1$
$Y = 0$	19404	168
$Y = 1$	208	31

On peut commencer par observer le pouvoir prédictif des modèles. Pour le premier modèle, les graves sont bien prédits dans 49% des cas et les non graves dans 76 % des cas. Parmi les sinistres qui sont estimés comme graves, 1.3% (Précision) sont effectivement graves contre 0.6% de manière générale. Le modèle fait deux fois mieux que le hasard.
Concernant le deuxième modèle, les graves sont bien prédits dans 13% des cas et les non graves dans 99 % des cas. Parmi les sinistres qui sont estimés comme graves, 16% sont effectivement graves contre 1.2% de manière générale. Le modèle fait beaucoup mieux que le hasard.

Cependant, il est possible de faire varier le seuil de classification en utilisant l'estimation de la probabilité d'être un grave. Si on veut augmenter le nombre de graves prédits, on peut prendre un seuil plus petit que 50% et au contraire si on veut faire baisser ce nombre, on peut prendre un seuil plus grand que 50%.
En reprenant les deux matrices, on peut se dire que pour le modèle de gauche, on pourrait faire baisser le nombre de graves prédits et au contraire, on pourrait un peu augmenter ce nombre pour le modèle de droite.

On va essayer de choisir le seuil optimal en le faisant varier et en observant la qualité du modèle. Pour cela, on va utiliser le score F1 et le score F2 introduits dans la partie 4.2.2.

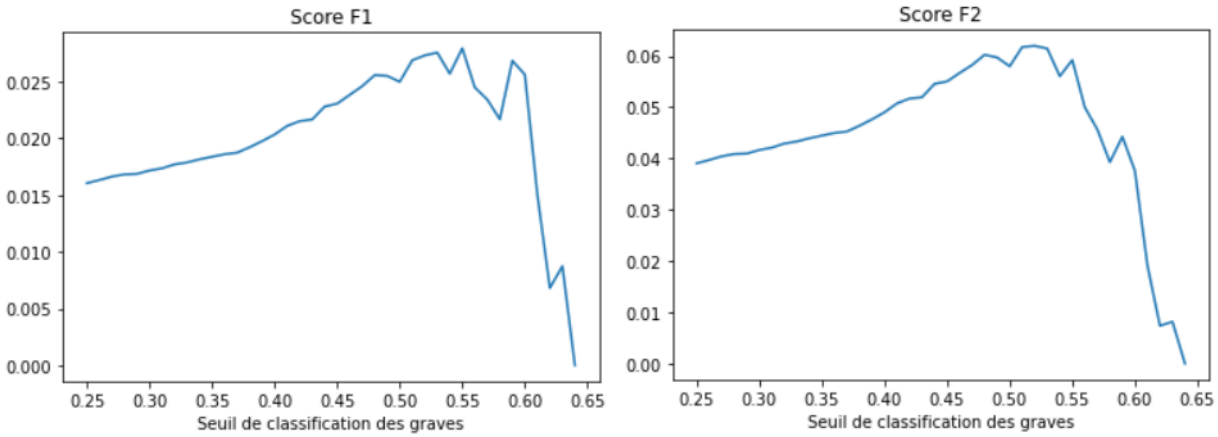

FIGURE 4.58 – Scores en fonction du seuil de classification pour le modèle en charge

Que l'on considère le score F1 ou le score F2, on constate que la zone de classification située entre 50% et 55% semble la plus performante. Comme pour ce modèle, on cherchait à pendre un seuil supérieur à 50%, **on peut décider de conserver un seuil à 55% dans la suite**.

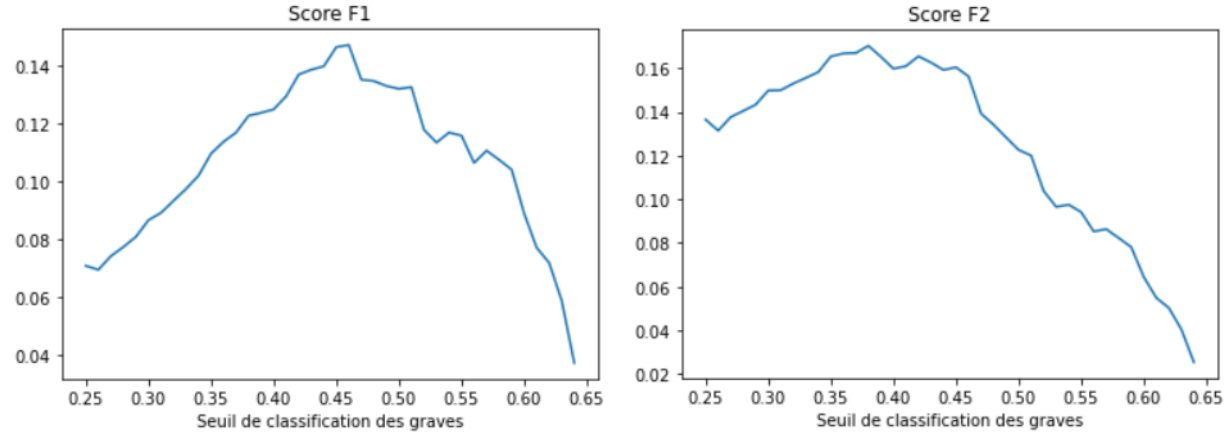

FIGURE 4.59 – Scores en fonction du seuil de classification pour le modèle en TD

Pour le modèle en taux de destruction, on constate cette fois-ci que la zone se situe plutôt entre 40% et 45%. **On décide de conserver la valeur**

de 45% dans la suite.

On obtient ainsi deux nouvelles matrices de confusion :

Charge	$f(X) = 0$	$f(X)) = 1$
$Y = 0$	17696	1987
$Y = 1$	98	30

TD	$f(X) = 0$	$f(X) = 1$
$Y = 0$	19294	278
$Y = 1$	198	41

Pour le modèle en charge, on est passé de 63 graves à 30 (division par 2.1). Le nombre de faux positifs est passé de 4723 à 1987 (division par 2.38). Le taux de grave passe ainsi de 1.3% à 1.5%. On a donc réussi à la fois à réduire le nombre de sinistres prédits comme grave et à légèrement augmenter la précision de la prédiction.
Pour le modèle en taux de destruction, le nombre de graves passe de 31 à 41 (augmentation de 32%) alors que les faux positifs augmentent de 168 à 278 (augmentation de 65%). Ici, aller chercher plus de sinistres graves a un coût : la précision baisse significativement (elle passe de 15.6% à 12.9%). Mais le rappel est passé de 13% à 17.2%, ce qui est positif.

Un des objectifs de cette modélisation étant la mise en place d'une politique de "dérisking" lors du renouvellement en fin d'année, on ne peut pas se permettre d'avoir un nombre de contrats à gérer trop grand. On cherche ainsi à avoir un nombre de sinistres prédits comme graves pas trop élevé. On va donc s'intéresser dans la suite au taux de graves en fonction du nombre de sinistres prédits comme graves.
C'est-à-dire qu'on applique notre modèle à la base de test puis on ne garde que les n sinistres ayant la plus forte probabilité d'être graves et on regarde le nombre de sinistres qui sont effectivement graves parmi ces n sinistres.

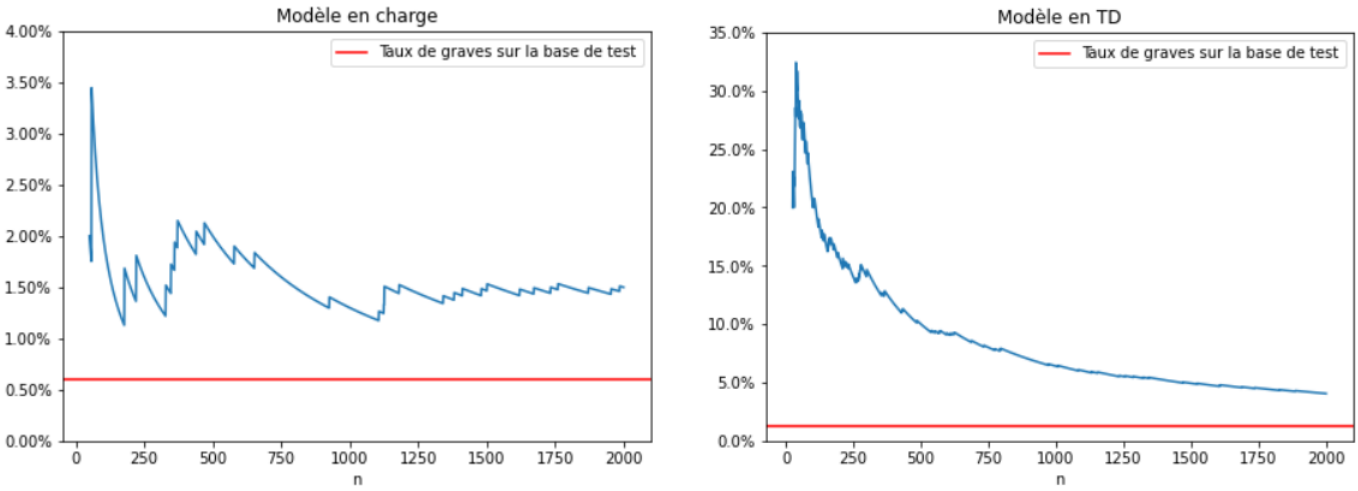

FIGURE 4.60 – Taux de Rappel parmi les n sinistres avec le score de propension le plus élevé

On constate que sur les deux graphes, on est clairement au-dessus de la droite rouge qui représente la proportion de sinistres graves sur l'ensemble de la base de test. Le modèle montre encore son caractère prédictif. Pour le modèle en charge, le rappel se stabilise entre 1.5% et 2%. Alors que pour le modèle en TD, le rappel est décroissant : on a intérêt à prendre un n assez petit puisque lorsqu'on augmente n, la proportion de graves baisse rapidement.

4.5 Gradient Boosting

4.5.1 Présentation

Le boosting[5] est une technique qui consiste en l'agrégation séquentielle de plusieurs modèles. L'idée est d'arriver au bon résultat en faisant des "petits pas" plutôt que d'essayer d'y arriver en une seule fois. De manière générale,les poids des individus sont modifiés dynamiquement et chaque modèle est pondéré selon sa performance.

Le Gradient Boosting[4] est une technique de boosting qui s'inspire de l'algorithme de descente de gradient en analyse réelle (Annexe E). Cependant, on cherche une fonction, non un point et on sait calculer le gradient que sur un échantillon limité de points (les observations de notre base). Dans le cas du Gradient Boosting, le modèle de base est l'arbre CART.
On nomme Φ notre fonction erreur que l'on cherche à minimiser. On commence par entrainer un premier arbre de type CART : f_0.

Ensuite, on va construire M arbres de cette manière :
Pour l'arbre m, on calcule l'opposée du gradient de l'arbre $m-1$ aux points d'observations, c'est-à-dire notre base d'entrainement :

$$r_{im} = \frac{-\partial \Phi(y,f)}{\partial f}\Big|_{\substack{y=y_i \\ f=f_{m-1}(X_i)}}$$

On entraine un arbre $\widetilde{f_m}$ sur la base $(X_1, r_{1m}), ..., (X_n, r_{nm})$.
On calcule $\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} \Phi\left(y_i, f_{m-1}(X_i) + \gamma \cdot \widetilde{f_m}(X_i)\right)$
On crée le nouveau modèle : $f_m(X_i) = f_{m-1}(X_i) + \lambda \cdot \gamma \cdot \widetilde{f_m}(X_i)$. C'est un modèle additif.
Le coefficient λ est appelé **coefficient de régularisation** ou de rétrécissement (shrinkage). Il est compris entre 0 et 1. Son but est de réduire le surapprentissage en ralentissant la convergence. Bien sûr, cela implique de devoir faire augmenter **le nombre M d'arbres dans le modèle**. En pratique cela se révèle être plus précis (mais plus coûteux en temps de calcul). M et λ sont deux hyperparamètres qui doivent être optimisés par validation croisée. Généralement il est conseillé d'avoir $\lambda < 0.1$ et $\frac{M}{\lambda} > 10$.
Dans le cas d'une classification binaire, on cherche avec f à estimer une probabilité. On prend généralement comme fonction Φ pour la classification la déviance binomiale :

$$\Phi(Y_i, f(X_i)) = -\mathbb{1}_{y_i=0} \cdot log\ \pi_0(X_i) - \mathbb{1}_{y_i=1} \cdot log\ \pi_1(X_i)$$

où $\pi_1(X_i) = \mathbb{P}\left[y_i = 1 | X_i\right] = \frac{exp(f_M(X_i))}{1 + exp(f_M(X_i))}$ et $\pi_0(X_i) = 1 - \pi_1(X_i)$
Ce qui donne :

$$\frac{-\partial \Phi(y,f)}{\partial f} = \mathbb{1}_{y_i=1} - \pi_1(X_i)$$

C'est donc l'écart entre la probabilité réelle d'appartenir à la classe 1 (qui vaut 0 ou 1) et la probabilité estimée. Si $y_i = 1$, on cherche à augmenter la probabilité d'autant plus qu'on est éloigné de 1 et si au contraire $y_i = 1$, on cherche à la réduire d'autant plus qu'on est éloigné de 0.

Il est possible pour encore plus améliorer la qualité du modèle en pratique de faire **du sous-échantillonnage (subsampling)** : lors de la construction de chaque arbre, seul un sous-échantillon de la base de données est utilisé

pour entrainer l'arbre. Ce sous-échantillon n'est pas un échantillon bootstrap : il est obtenu par tirage sans remise. On parle de **stochastic gradient boosting**. Généralement, on en profite pour restreindre le nombre de variables à considérer lors de la construction d'un nouveau nœud de la même manière que pour un random forest.

Pour pénaliser des modèles trop complexes, il est possible d'introduire une pénalisation lorsqu'on utilise une variable pour la première fois. Mathématiquement, on modifie le gain en hétérogénéité G d'un split $X_j \leq s$:

$$\widetilde{G} = G - \alpha_1 \cdot U(X_j) - \alpha_2 \cdot V(X_j)$$

où $\alpha_1 \geq 0$ et $\alpha_2 \geq 0$ sont les deux coefficients de régularisation.
où $U(X_j)$ vaut 1 si la variable X_j n'a pas encore été utilisée dans l'arbre en cours de construction et 0 sinon.
où $V(X_j)$ vaut 1 si la variable X_j n'a pas encore été utilisée dans la forêt en construction et 0 sinon.

C'est une **pénalisation par variable**.

4.5.2 Application

Après l'application des méthodes d'optimisation des paramètres présentées, voici les principaux paramètres des Stochastic Gradient Boosting obtenus :

	Modèle en charge	Modèle en TD
Nombre d'arbres	500	500
Coefficient de régularisation	0.001	0.002
Indice d'hétérogénéité	Entropie	Entropie
Subsampling	80%	100%
Variables retenues	70%	70%
Pondération des graves	équilibrée	équilibrée
Profondeur maximale	10	15
Nombre maximal de noeuds	25	27

TABLE 4.12 – Les principaux hyperparamètres

Comme précédemment, on fixe le seuil de classification à 50% et on obtient ces matrices de confusion :

Charge	$f(X) = 0$	$f(X)) = 1$
$Y = 0$	16996	2687
$Y = 1$	84	44

TD	$f(X) = 0$	$f(X) = 1$
$Y = 0$	16316	3256
$Y = 1$	112	127

On observe encore une fois le pouvoir prédictif du modèle, on a une précision de 1.6% (0.6% pour le hasard) et 15.6% (1.2% pour le hasard). Les graves sont bien prédits dans quasiment 50% des cas pour le modèle en charge et dans 13% des cas dans le modèle en taux de destruction.

On étudie l'évolution du score F1 et du score F2 en fonction du seuil de classification.

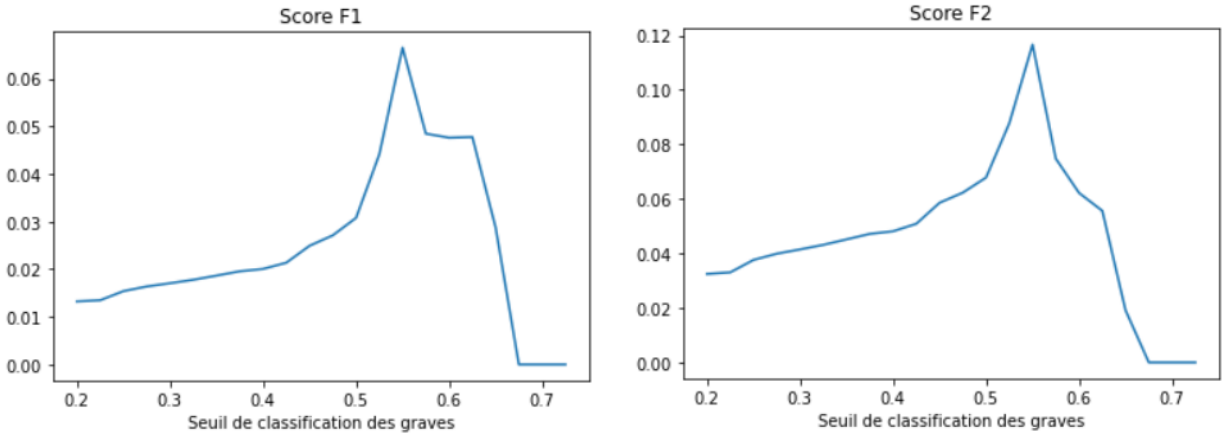

FIGURE 4.61 – Scores en fonction du seuil de classification pour le modèle en charge

Que l'on regarde le score F1 ou le score F2, on constate que la zone de classification située aux alentours des **55**% semble optimale.

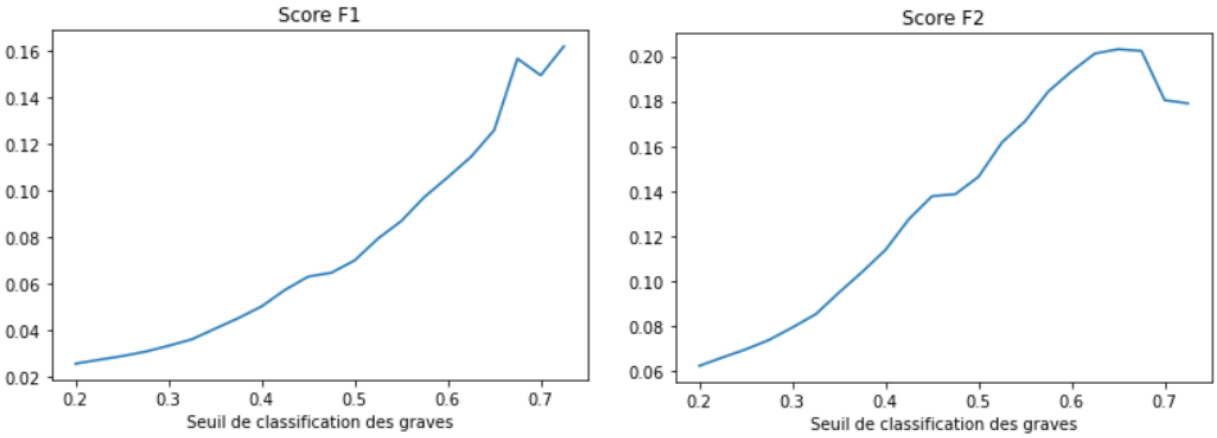

FIGURE 4.62 – Scores en fonction du seuil de classification pour le modèle en TD

Pour le modèle en taux de destruction la zone semble plutôt se situer entre 60% et 70%. **On décide de conserver la valeur de 65% dans la suite**.

On obtient ainsi deux nouvelles matrices de confusion :

Charge	$f(X) = 0$	$f(X)) = 1$
$Y = 0$	18939	744
$Y = 1$	98	30

TD	$f(X) = 0$	$f(X) = 1$
$Y = 0$	18591	981
$Y = 1$	157	82

Les résultats sont clairement meilleurs, pour le modèle en charge, on a divisé par 3.6 le nombre de faux positifs alors que le nombre de vrais positifs n'a été divisé que par 1.5. Pour le modèle en taux de destruction, on a divisé par 3.3 le nombre de faux positifs tandis que le nombre de vrais positifs n'a été divisé que par 1.5.

On regarde encore une fois les n sinistres avec le score de propension le plus élevé.

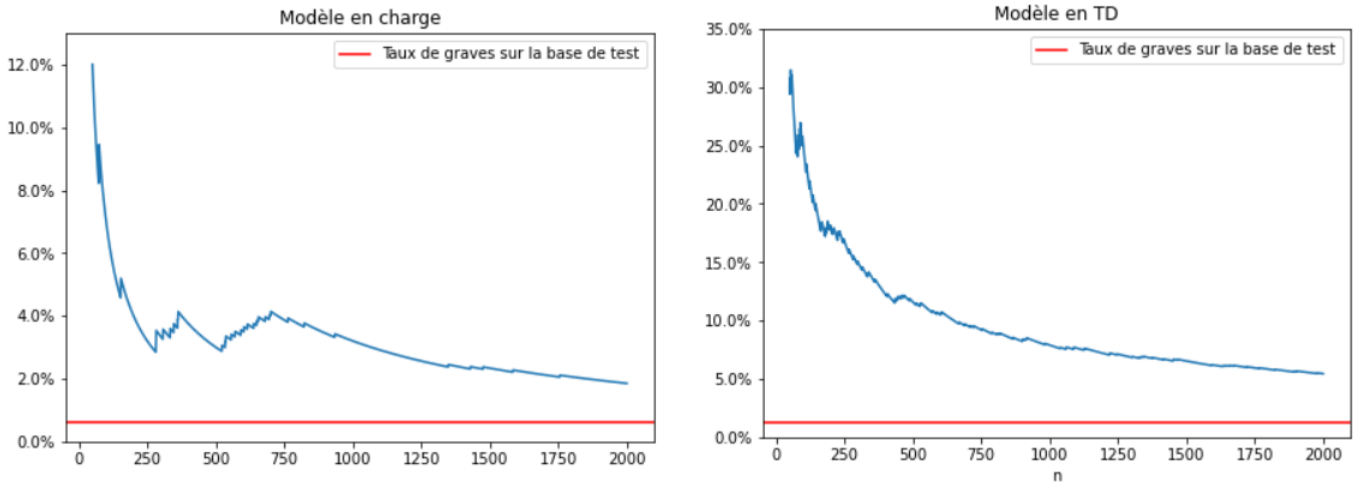

FIGURE 4.63 – Taux de Rappel parmi les n sinistres avec le score de propension le plus élevé

Comme pour le modèle Random Forest, les sinistres classés comme les plus à risque ont clairement plus tendance à être graves. Pour le modèle en charge, on a intérêt à prendre n proche de 750 et pour le modèle en TD, un n inférieur à 500.

Analyse des résultats

Dans cette dernière partie, on commence par comparer les deux modèles utilisés dans la partie précédente : Random Forest et Gradient Boosting. Ensuite, on essaye de comprendre l'importance et le pouvoir prédictif des différentes variables explicatives. Enfin, on compare les deux approches : celle par taux de destruction et celle par charge.

5.1 Comparaison des modèles

On va essayer de déterminer si un modèle est plus performant que l'autre. Pour rappel,on a déterminé pour chaque modèle un seuil de classification "optimal". On peut donc commencer par comparer les scores F1 et F2 associés :

Charge	Random Forest	Gradient Boosting
F1	0.028	0.066
F2	0.064	0.117

TD	Random Forest	Gradient Boosting
F1	0.14	0.13
F2	0.15	0.20

Pour le modèle avec le seuil à 100 000€, le modèle utilisant le Gradient Boosting est beaucoup plus performant. On cherche à maximiser les scores F1 et F2. Et on constate que le score F1 est 2.4 fois plus élevé et que le score F2 est 1.8 fois plus élevé. Selon ces deux critères, le modèle Gradient Boosting est significativement meilleur.
Pour le modèle avec le seuil à 12%, le score F1 ne varie pas beaucoup selon le type de modèle utilisé. Mais, concernant le score F2, le Gradient Boosting est légèrement supérieur avec un score 33 % plus élevé. On constate qu'ici aussi le Gradient Boosting semble plus performant, mais que l'écart est moins significatif.

Concernant l'AUC, pour le modèle en charge, le Random Forest et le Gradient Boosting donnent tous les deux un AUC de 71%. Pour le modèle

en taux de destruction, l'AUC du Random Forest est de 72% tandis qu'il est de 77% pour le Gradient Boosting. l'AUC permet seulement de départager les modèles dans l'approche par taux de destruction en donnant un avantage au Gradient Boosting.

On va maintenant comparer le taux de gravité parmi les sinistres avec le score de propension le plus élevé pour chaque modèle. On regarde ainsi le taux de gravité aux alentours de n proche de 250, 500 et 750 :

Charge	Random Forest	Gradient Boosting
250	1.5%	3%
500	2%	3%
750	1.5%	4%

TD	Random Forest	Gradient Boosting
250	15%	17%
500	10%	12%
750	8%	10%

Pour le modèle en charge, le Gradient Boosting se démarque encore une fois comme étant le modèle avec le pouvoir prédictif le plus important (de 50% à 150% plus performant). Pour le deuxième modèle, le Gradient Boosting est légèrement plus performant (de 13% à 25% plus performant).

Enfin, on va utiliser le κ de Cohen pour comparer les modèles :

κ	Random Forest	Gradient Boosting
Charge	0.064	0.075
TD	0.153	0.174

On peut déjà constater que les valeurs sont toutes positives, ce qui est le minimum attendu. Mais les valeurs sont très faibles par rapport à ce que l'on pourrait attendre. On peut expliquer cela par le fait que les classes sont déséquilibrées. On peut toutefois raisonner en valeur relative. On confirme encore une fois que le Gradient Boosting fournit un meilleur modèle. Dans les deux cas, la valeur du κ de Cohen est supérieure.

En conclusion, on peut affirmer que **le Gradient Boosting est de meilleure qualité** que ça soit pour le seuil à 100 000€ ou le seuil à 12%.

5.2 Importance des variables

Généralement, les méthodes utilisées sont des méthodes paramétriques. On peut citer les méthodes GLM. Avec ces méthodes, il est facile de comprendre l'importance et l'influence des variables, car ces méthodes fournissent des coefficients explicites qui sont interprétables. Lors de l'utilisation d'une méthode non paramétrique, il n'est pas possible d'utiliser ce type de raisonnement : ceux sont des "boites noires". Cependant, pour les méthodes CART, il existe des moyens d'en apprendre un peu plus sur l'importance et l'influence des variables.

5.2.1 Importance d'une variable dans une méthode CART

Comme on l'a vu en partie 4.3, la construction d'un arbre est basée sur la notion d'hétérogénéité. Notamment, pour la construction d'un nouveau nœud, on cherche à maximiser le gain d'hétérogénéité. On va assimiler l'importance d'une variable par son pouvoir à faire croître l'hétérogénéité.

Pour la variable d'indice i, on définit Ω_i l'ensemble des nœuds de l'arbre qui utilisent la variable i pour son split. Soit G_j le gain en hétérogénéité du nœud j. On définit alors la valeur d'importance de la variable i par :

$$\theta_i = \sum_{j \in \Omega_i} G_j$$

θ_i est le gain en hétérogénéité qui revient à la variable i. Pour obtenir la proportion de gain en hétérogénéité qui lui revient, on divise cette valeur par le gain total :

$$\omega_i = \frac{\theta_i}{\sum_{j=1}^{k} \theta_j}$$

De cette manière, on peut hiérarchiser l'ensemble des variables explicatives par leur pouvoir discriminant.

Cependant, cette méthode possède de nombreuses faiblesses. D'abord, on connait uniquement l'importance des variables et par conséquent on ne connait pas leur influence. C'est-à-dire que l'on peut savoir si une variable est importante ou non, mais on ne sait pas si cette variable a tendance à faire augmenter la probabilité d'avoir un grave ou au contraire la faire baisser.

De plus, l'importance est calculée à partir de la base d'entrainement (qui a servi à construire l'arbre). Cette méthode explique l'importance au sein de la base d'entrainement. Pour pouvoir vraiment interpréter les résultats de manière prédictive, il faut que le modèle soit performant lors de la prédiction et que le surapprentissage soit limité. Enfin pour une variable catégorielle avec beaucoup de modalités, le résultat peut être faussé du fait que cette variable est divisée en de nombreuses variables binaires. L'importance d'une telle variable étant obtenue en sommant l'importance des sous variables binaires, l'importance peut être anormalement élevée, car ce type de variable peut favoriser le surapprentissage. Une variable qui favorise le surapprentissage aura une importance élevée dans le modèle, mais d'un point de vue prédictif, cette variable n'a que très peu d'importance. Ainsi, pour pouvoir utiliser cette donnée, il est impératif de limiter le surapprentissage.

On peut facilement généraliser l'importance des variables dans un arbre au cas d'un modèle composé de plusieurs arbres en sommant les gains en hétérogénéité G_j sur l'ensemble des arbres plutôt qu'un seul pour obtenir θ_i. Dans le cas où la forêt introduit une pondération, il suffit de pondérer de la même manière les θ_i.

Voici les dix variables les plus importantes pour les deux modèles Gradient Boosting retenus :

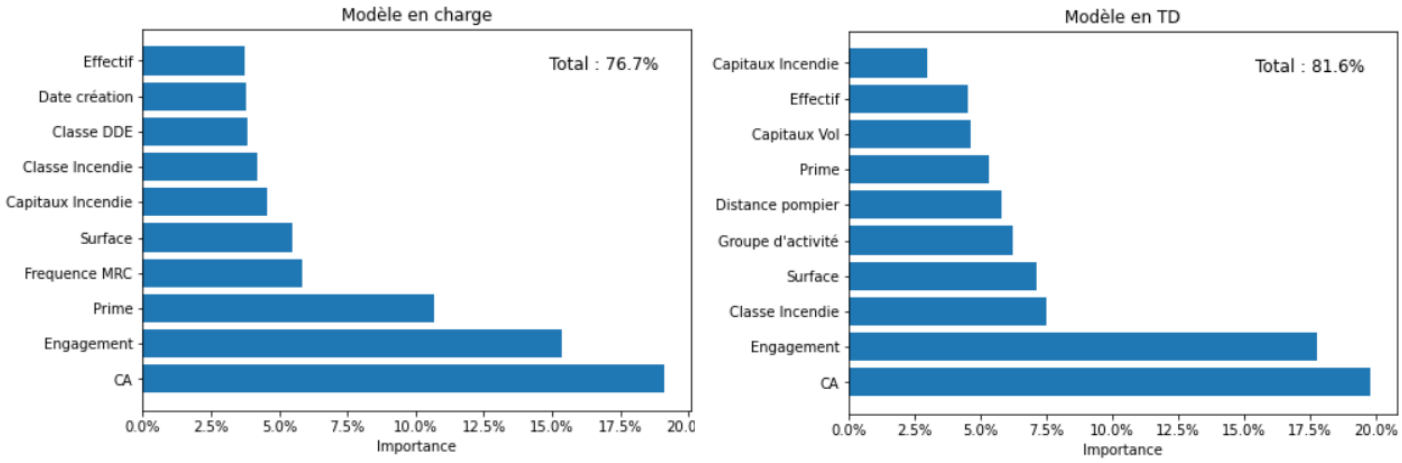

FIGURE 5.64 – 10 variables les plus importantes pour les modèles Gradient Boosting

On constate l'importance significative du chiffre d'affaires et de l'engage-

ment dans les deux modèles avec entre 35% et 40% du gain en hétérogénéité expliqué par ces deux seules variables. La prime et la surface sont aussi très importantes dans les deux modèles avec une importance supérieure à 5%.

La sinistralité passée en MRC est importante (5.8%) dans le modèle en charge mais peu significative (1%) dans le modèle en TD. La date de création de l'entreprise a une importance de 3.8% dans le modèle en charge contre 2.6% dans le modèle en TD. La classe Dégâts des Eaux a une importance de 3.8% dans le premier modèle contre 1.6% dans le deuxième.

L'effectif a une importance similaire dans les deux modèles : 3.6% dans le premier et 4.5% dans le deuxième.

Pour le modèle en TD, on constate un pouvoir discriminant plus important du groupe d'activité (6.2% contre 0.6%) et de la distance aux pompiers (5.8% contre 1.1%).

Cette différence se comprend en partie avec l'analyse faite dans la partie 2.5.1. On y constate une plus forte prédominance des sinistres Incendie et Bris de Machines dans les sinistres graves en charge tandis que les sinistres graves en TD offrent une plus grande place au vandalisme et aux dégâts des eaux.

Il y a une certaine concentration de l'importance des variables. Il y a un peu plus de 50 variables explicatives et les 10 plus importantes pèsent 76% dans le modèle en charge et 81% dans le modèle en TD. De nombreuses variables ne sont quasiment pas utilisées. On peut par exemple citer le code territoire (Province/Ile de France) qui pèse pour moins de 1%, la date de construction du bâtiment qui pèse elle aussi moins de 1%. Les variables construites avec la sinistralité antérieure sont quasiment toutes (sauf la sinistralité MRC : S_FREQ_MRC) peu utilisées.

Pour rappel, un des objectifs de l'étude était d'étudier l'importance de certaines variables externes complètement nouvelles. Voici un tableau qui présente l'importance des variables externes :

	Charge	TD
Note N-1	0.46%	0.37%
Note N	1.13%	1.49%
Variation Note	2.40%	0.61%
Distance pompiers	1.15%	5.78%
Prix m^2	0.37%	1.04%
Année construction	0.15%	0.49%
Date de création	3.78%	2.62%

Le prix du m^2 et l'ancienneté du bâtiment ne sont que très peu significatives. On peut ajouter que ces variables présentent un taux de vides assez élevé. La date de création de l'entreprise et la distance aux pompiers ont une importance plutôt élevée comme dit précédemment. Les notes financières jouent un léger rôle dans le modèle en charge (3.99%) mais sont moins importantes dans le modèle en taux de destruction (2.47%). Il semble que cela soit surtout la variation et la note à l'année N qui comptent.

5.2.2 Coefficient SHAP

La valeur de Shapley est issue de la théorie des jeux. Lors d'un jeu collaboratif, cette valeur sert à répartir le gain d'un jeu entre les joueurs. Dans le cas d'une classification, on utilise le coefficient SHAP (Shapley Additive Explanation) [8]. Si on a k variables explicatives, la méthode consiste en la décomposition de la prédiction de cette manière :

$$f(X_i) - \mathbb{E}[f(X)] = \sum_{j=1}^{k} \phi_{j,i}$$

C'est-à-dire que l'écart de la prédiction à la moyenne est expliqué par une somme de contributions de chaque variable explicative. Cette contribution peut être positive ou négative. Ce coefficient SHAP est intéressant, car il permet de quantifier l'importance d'une variable ($|\phi_{j,i}|$) tout en lui affectant son influence (le signe de $\phi_{j,i}$).

Pour calculer $\phi_{j,i}$, on commence par définir Σ_j l'ensemble des permutations de $[\![1, k]\!] \setminus \{j\}$. Puis on calcule :

$$\phi_{j,i} = \sum_{S \subseteq \Sigma_j} \frac{|S|! \cdot (k - |S| - 1)!}{k!} \left(\widetilde{f}(X_{S \cup \{j\}}) - \widetilde{f}(X_S) \right)$$

où $\widetilde{f}(X_\Omega)$ représente la prédiction du modèle à partir des variables dans Ω (ensemble d'indices de variables explicatives) et $|S|$ le cardinal de l'ensemble S d'indices.
Le terme $\widetilde{f}(X_{S \cup \{j\}}) - \widetilde{f}(X_S)$ représente la différence de prédiction entre le modèle où l'on utilise les variables de S et la variable j et le modèle où l'on utilise les mêmes variables sauf j. **Ce coefficient représente ainsi l'apport marginal de la variable j toutes choses égales par ailleurs.** Dans le cas d'une classification binaire, ce terme vaut 0 si la prédiction est la même (variable "inutile") et ± 1 si la prédiction est différente.
Le terme $\frac{|S|! \cdot (k - |S| - 1)!}{k!}$ représente la probabilité d'obtenir le tirage S (probabilité de la permutation dans l'ensemble des permutations Σ_j).
Le terme $\phi_{j,i}$ s'interprète ainsi comme l'espérance des effets marginaux de la variable j pour la prédiction de X_i.

Une fois avoir calculé l'ensemble des $\phi_{j,i}$ pour toutes les variables pour tous les individus, on peut présenter les résultats sous forme graphique :

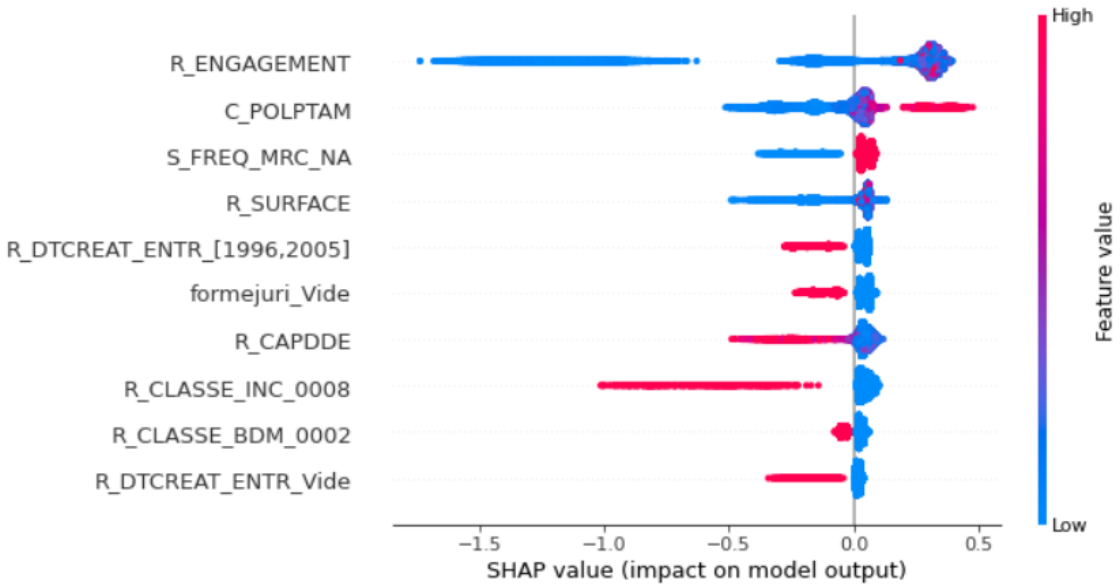

FIGURE 5.65 – Valeurs SHAP pour les 10 variables les plus importantes (modèle en charge)

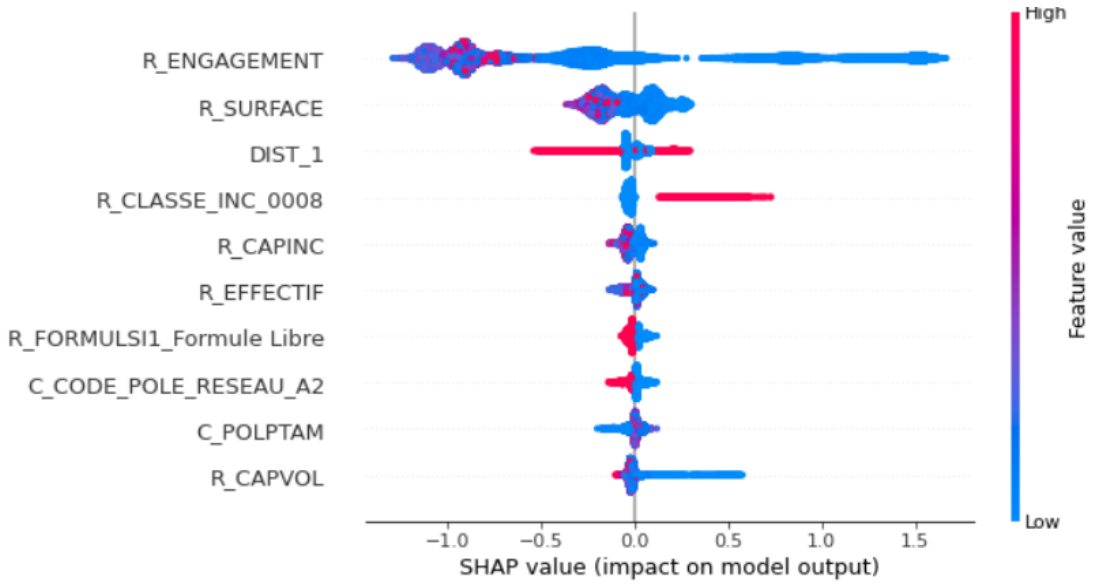

FIGURE 5.66 – Valeurs SHAP pour les 10 variables les plus importantes (modèle en TD)

Sur le graphique, une ligne correspond à une variable. Chaque variable est classée selon son importance de manière décroissante. L'importance d'une variable j est mesurée en calculant $|\phi_j| = \sum_{i=1}^{n} \frac{|\phi_{j,i}|}{n}$. Sur chaque ligne, les points correspondent à l'ensemble des individus de la base de test. Ces points sont placés selon sa valeur SHAP ($\phi_{j,i}$) en abscisse. L'épaisseur de la barre représente la densité de point. Ainsi les zones où la barre s'épaissit correspondent à une zone avec beaucoup d'individus qui ont des $\phi_{j,i}$ proches. Enfin, la couleur des points représente la valeur de la variable : quand le point est plutôt rouge alors la valeur de la variable associée pour cet individu est clairement supérieure à la valeur moyenne de cette variable. Au contraire, quand le point est bleu, la valeur de la variable associée pour l'individu est très inférieure à la valeur moyenne de cette variable. Pour les variables binaires, le bleu correspond à 0 et le rouge à 1.

On peut faire la remarque que cette méthode donne l'importance et l'influence de la variable lors de la prédiction. En effet, on utilise l'impact de la variable lors des prédictions sur la base de test pour obtenir les $\phi_{j,i}$. Cette méthode semble ainsi plus fiable que la méthode précédente. Cependant, elle requiert un temps de calcul très important, car pour chaque individu de la base de test, on utilise deux modèles par variables. On peut ajouter que ce coefficient n'est pas fiable en cas de faible effectif pour une variable catégo-

rielle, car on a peu de points pour estimer la moyenne.

Ce type de graphique peut paraitre illisible quand le nombre de points est trop important. On peut synthétiser l'information en étudiant la corrélation entre la variable et le coefficient de SHAP. C'est ce qu'on représente sur les graphiques suivants (où le rouge correspond à une corrélation positive et le bleu une corrélation négative) :

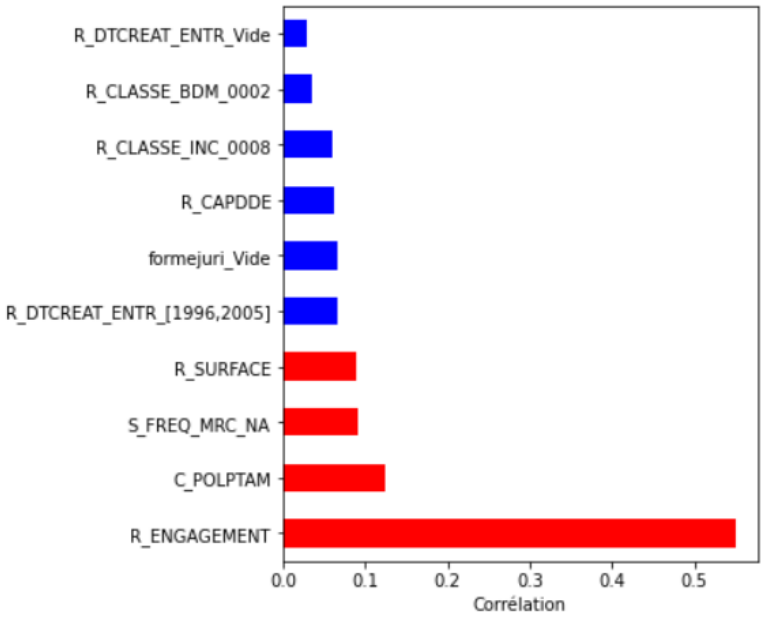

FIGURE 5.67 – Corrélation entre la valeur SHAP pour les 10 variables les plus importantes (modèle en charge)

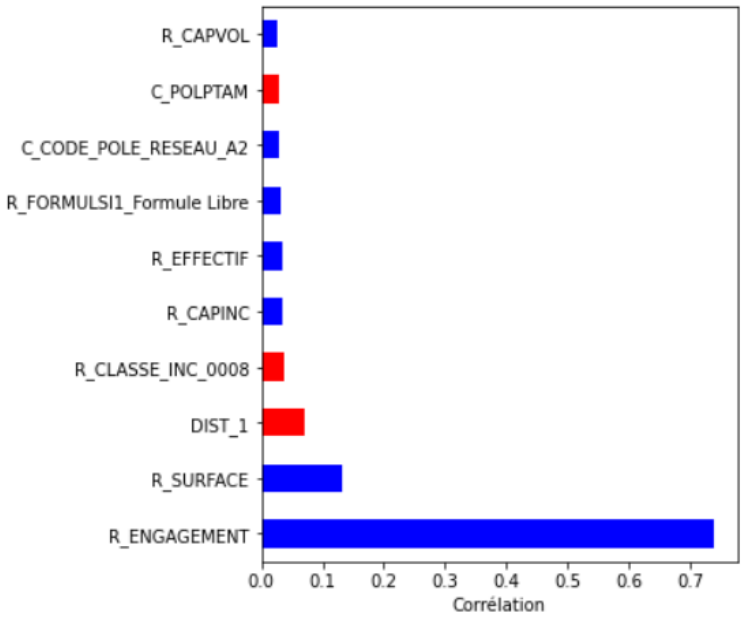

FIGURE 5.68 – Corrélation entre la valeur SHAP pour les 10 variables les plus importantes (modèle en TD)

Déjà, on observe une certaine cohérence entre les résultats ci-dessus et les résultats obtenus dans la partie précédente concernant l'importance des variables pour chaque modèle : les variables qui ressortent sont les mêmes. On remarque cependant une grande différence concernant le chiffre d'affaires qui semble beaucoup moins discriminant selon la méthode SHAP. Il ressortait comme étant le plus important dans les deux modèles avec la méthode précédente. Ici, il n'apparait qu'à la 16e place (modèle en charge) et 13e place (modèle en TD). On peut faire la remarque que la variable $S\_FREQ\_MRC\_NA$ correspond à la variable binaire représentent les affaires nouvelles.

Quand on essaye d'analyser le sens de l'influence de chaque variable, on observe une différence entre les deux modèles. Les variables qui caractérisent la taille de l'entreprise et la taille potentielle du risque associé (engagement, prime, surface, capitaux ...) ont clairement une influence positive sur le risque d'avoir un grave en termes de charge alors qu'ils ont une influence négative sur le risque d'avoir un grave en termes de TD. On identifie ainsi clairement la taille de l'entreprise comme un facteur de risque déterminant pour l'approche par charge et au contraire, pour l'approche par taux de destruction, la taille de l'entreprise a tendance à réduire le risque. On peut illustrer cette différence en prenant l'exemple de l'engagement et de la prime avec ces nuages de

points :

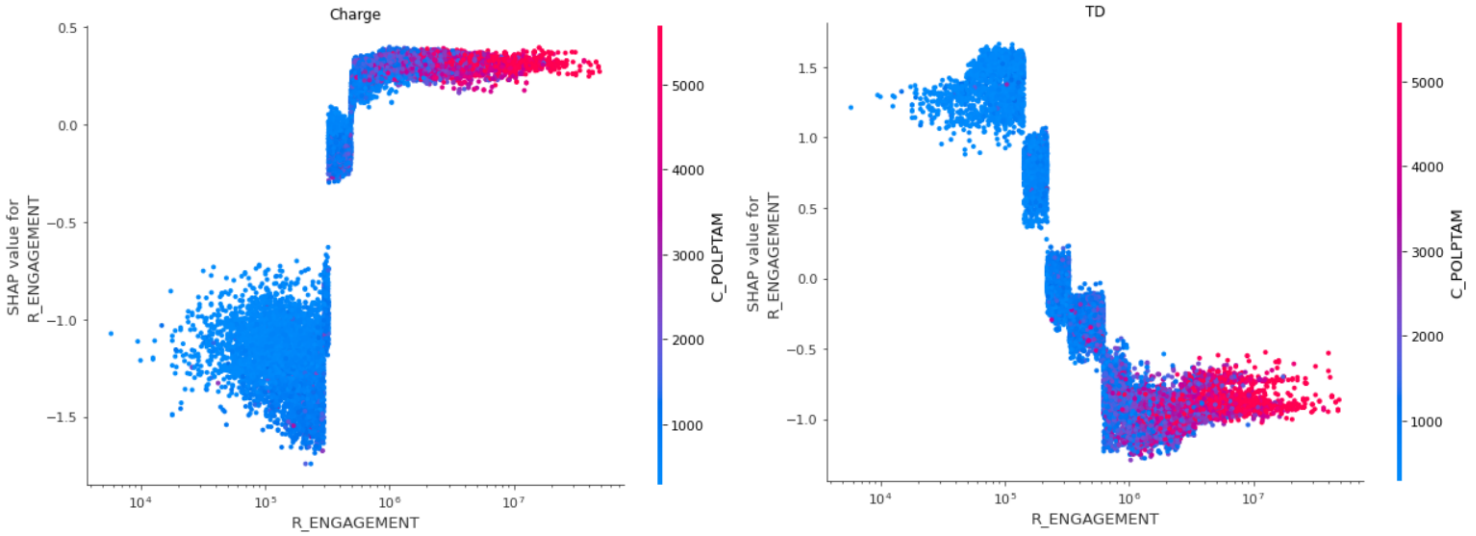

FIGURE 5.69 – Valeur SHAP en fonction de l'engagement, échelle logarithmique en abscisse

Les couleurs des points représentent les valeurs de la prime. Les points rouges représentent les primes les plus élevées et les bleus les plus faibles. Sur le nuage de points de gauche, on constate clairement que la valeur de SHAP augmente avec l'engagement et la prime. Sur le nuage de droite, c'est l'inverse, les valeurs de SHAP associées à des engagements élevés et des primes élevées sont les plus faibles. On observe des paliers sur les deux graphiques, cela vient du fait que les algorithmes sont basés sur des arbres CART (pas linéaire). Les paliers correspondent sans doute à des seuils de split utilisés dans les arbres les plus importants.

Pour les sinistres supérieurs à 100000€, on peut citer quelques facteurs aggravants ou atténuants (en dehors des variables relatives à la taille) :

Facteur aggravant	Facteur atténuant
Affaire nouvelle	Classe Incendie 008
Enteprise créée après 1995	Classe BDM 002
Distance aux pompiers élevée	Classe Incendie 001
Variation de la note financière de -2 ou -3	Code naf de type 8
Forme juridique de type 9	Entreprise créée avant 1995
Ne pas avoir de protection antivol	Zone BDM V13
Avoir le produit 100% Artisans-Commerçants	Zone Incendie V02

La forme juridique 9 rassemble principalement des associations. Ici, les associations semblent avoir un risque plus élevé. Cependant, les associations représentent une très faible minorité du portefeuille et il se peut qu'il y ait un léger phénomène de sur apprentissage.
Les notes financières n'ont pas un grand impact selon le coefficient de SHAP mais on arrive quand même à identifier un risque aggravant quand il y a une dégradation de la note.
Un des facteurs qui semble ressortir est l'ancienneté de l'entreprise. En effet, les entreprises récentes semblent avoir un risque plus élevé que les entreprises plus âgées.
Une distance aux pompiers assez élevée semble aussi avoir un léger effet aggravant. Ce résultat est un résultat qui était attendu.
Dans les facteurs aggravants, il y a aussi le fait d'avoir choisi le produit destiné particulièrement aux artisans. De manière générale, le fait d'être une entreprise avec une activité artisanale est aggravant.
On constate aussi qu'être dans la classe Incendie 008 est un facteur atténuant. Or cette classe est la classe identifiée comme étant la plus risquée lors de la tarification. Ce résultat peut paraître paradoxal. Pour l'expliquer, on peut commencer par dire que les représentants de cette classe sont peu nombreux. De plus, on peut penser que le risque est bien identifié pour ces entreprises et donc que des mesures de préventions adéquates sont appliquées.
Le code NAF de type 8 correspond principalement à des activités de bureau et de service. Il est attendu que les entreprises ayant ce type d'activité aient un risque plus faible.

On peut vérifier l'estimation du score de propension moyen en bivarié pour la classe de risque Incendie :

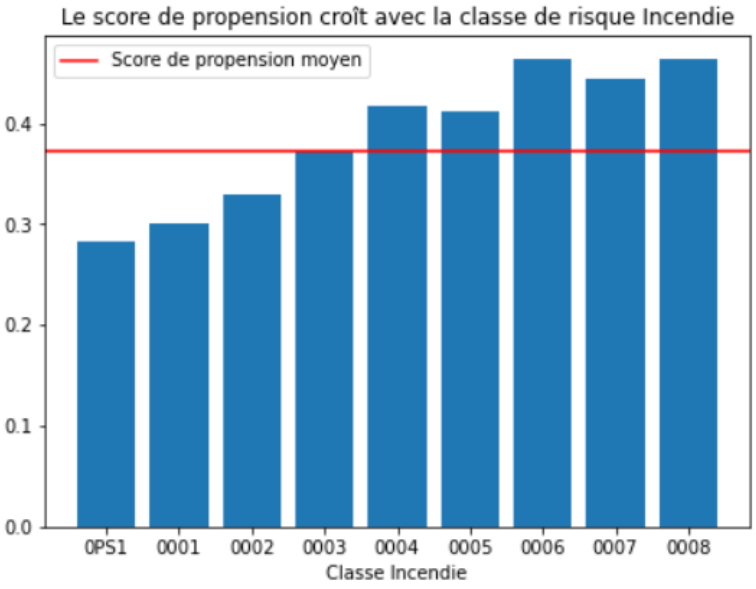

FIGURE 5.70

On observe bien que le score de propension est d'autant plus élevé que le risque incendie l'est aussi. Ainsi le fait que la classe 0008 soit identifiée comme facteur atténuant n'implique pas que le risque de cette classe soit plus faible. L'information du risque est déjà contenue dans d'autres variables. Le coefficient SHAP s'interprète bien toutes choses égales par ailleurs.

Pour les sinistres avec un TD supérieur à 12%, voici quelques facteurs aggravants ou atténuants (en dehors des variables relatives à la taille) :

Facteur aggravant	Facteur atténuant
Classe Incendie 006, 007 et 008	Avoir choisi la formule libre
Note financière mauvaise (année N)	Réseau courtier
Entreprise récente	Classe Incendie 001 ou 002
Distance aux pompiers faible	Être en province
État financier qui s'est détérioré (-2 ou -3)	Avoir une protection mécanique (antivol)
Entreprise de type Hôtel	Être un prestataire de services
Ne pas avoir de protection antivol	Entreprise créée avant 1995
Être en région parisienne	Classe DDE 001 ou 002

Les classes Incendies 006, 007 et 008 correspondent aux classes de risque les plus élevées et les classes 001 et 002, celles avec le risque le plus faible. Ici, la différence concernant les classes à haut risque et les classes à faible risque est bien observable. Les classes de risque les plus élevées sont des facteurs aggravants et les classes les plus faibles sont des facteurs de risque atténuants. La note financière a un impact faible, mais dont le sens est clair. Le fait d'avoir un état financier de mauvaise qualité ou avoir connu une dégradation de son état financier est un facteur aggravant pour ce type de sinistres graves. Concernant l'activité de l'entreprise. Les hôtels ressortent comme étant plus à risque alors que les prestataires de services sont encore identifiés comme moins risqués.
On constate aussi une opposition entre la région parisienne et la province. On peut peut-être rapprocher cela au fait qu'être proche des pompiers est un facteur de risque aggravant. En effet, une distance très faible aux pompiers signifie souvent que l'on est localisé dans un milieu urbain. Ainsi on pourrait en déduire que les entreprises situées dans les grandes agglomérations (type région parisienne) sont plus sujettes à des sinistres graves en termes de taux de destruction. Ce qui pourrait invalider cela est le fait que le prix du m^2 ne semble pas du tout être une variable ayant un impact.
Enfin, la protection antivol intervient encore ici. Ne pas avoir de protection semble être un facteur aggravant alors qu'avoir un système de protection mécanique est un facteur qui atténue le risque de grave.

Concernant les notes financières, une analyse bivariée du score de propension donne un résultat intéressant :

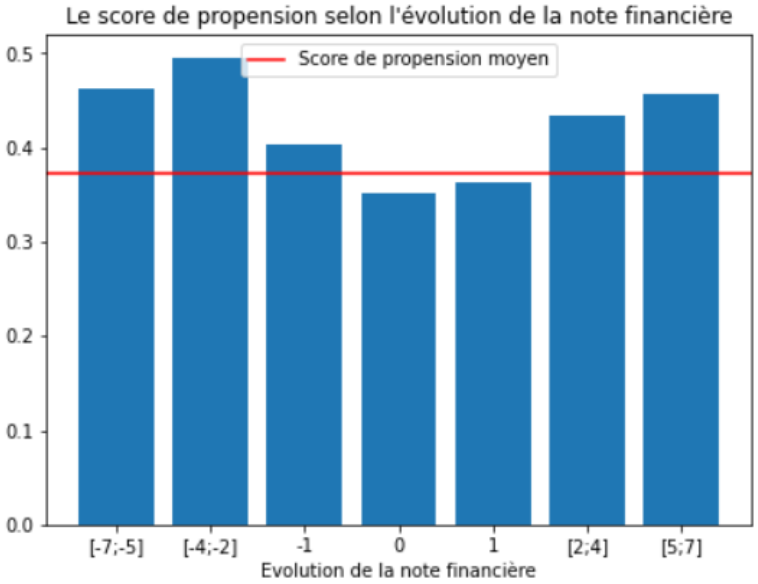

FIGURE 5.71

On constate que le score semble beaucoup plus élevé pour les entreprises qui ont connu une évolution significative (positive comme négative). Cependant, il faut garder en tête que ces entreprises sont peu nombreuses. Pour la grande majorité des entreprises, la variation est faible ou on n'a pas la note financière à disposition.

5.3 Comparaison des deux approches

Sur la base des deux modèles Gradient Boosting, on a clairement identifié de nombreuses différences entre les deux approches. L'approche par taux de destruction fait ressortir des risques associés à des petites entreprises. Il est probable que ces petites entreprises (récentes) soient moins matures financièrement. Ainsi, un état financier plutôt mauvais ou qui se dégrade augmente la chance d'être grave. Ce phénomène est amplifié quand l'entreprise est en région parisienne et quand l'activité implique un risque Incendie plus élevé. L'approche par charge du sinistre fait ressortir des risques liés aux grandes entreprises. Ce risque est amplifié quand l'entreprise est récente ou quand l'état financier de l'entreprise s'est dégradé. Les risques identifiés semblent souvent liés aux sinistres Incendie ou BDM. Les activités classées comme plus à risque concernant les incendies ou les bris de machines ont un risque plus

élevé.
Dans les deux approches, les prestataires de services et les entreprises plutôt anciennes (création avant 1995) ont un risque atténué. La dégradation récente de la note financière est un risque aggravant dans les deux approches.

Du point de vue de la qualité du modèle, on a vu que le modèle utilisant le taux de destruction est légèrement plus performant. Pour rappel, on a une AUC de 77% pour le modèle en TD (contre 71%). De manière générale, la détection des graves semble un peu plus simple dans le modèle par taux de destruction. Dans les deux cas, on peut adapter la précision et le rappel en faisant varier le seuil de classification. Il est fort probable que l'approche par TD semble un peu plus performante, car il y a un peu plus de sinistres graves avec cette approche. Comme on a plus de données sur les sinistres graves, le modèle arrive mieux à discriminer par la suite.

D'un point de vue opérationnel, il semble que cela sera le modèle par charge qui sera retenue. En effet, ce qui est recherché en priorité est la réduction de la charge. Il est préférable de prévenir quelques sinistres coutant plusieurs centaines de milliers d'euros plutôt que des sinistres en plus grands nombres, mais coutant beaucoup moins.

Conclusion

On a d'abord déterminé deux seuils de gravité selon deux approches différentes en utilisant la Théorie des Valeurs Extrêmes. Une première approche classique utilisant la charge du sinistre avec un seuil à 100 000€. Une seconde approche utilisant le taux de destruction fournit un seuil à 12%. Ces deux approches fournissent une modélisation différente, car on a remarqué que les sinistres concernés n'étaient pas les mêmes. L'approche utilisant le seuil 100 000€ semble le plus correspondre à ce que recherche Generali.

Ensuite, on a essayé de modéliser la propension d'un sinistre à être grave avec des méthodes de Machine Learning. Dans les deux approches, le meilleur modèle est un modèle de type Gradient Boosting.

Enfin, une analyse des modèles retenus a permis d'en apprendre plus sur l'effet des différentes variables explicatives. L'effet de beaucoup de variables est moins important qu'attendu notamment à cause du taux de complétion faible ou du peu de précision.

L'objectif initial étant d'avoir une meilleure compréhension de la sinistralité grave en MRC. Ces résultats, combinés avec les résultats du reste de l'équipe LCAP qui modélise la distribution du nombre de sinistres, sont dans l'ensemble cohérents avec la surveillance actuelle du portefeuille. Pour améliorer la gestion du risque grave, parmi les mesures envisagées, on peut citer une amélioration de la prise en compte de certaines variables qui sont importantes dans la modélisation des graves et une revue de certaines polices qui vont devoir repasser par la souscription.

Bibliographie

[1] L. Breiman (2001). Random forests. *Machine Learning, vol.45, no 1, p. 5–32.*

[2] A.L.M. Dekkers, J.H.J. Einmahl, and L. DeHaan (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Stat., 17(4), 1833-1855.*

[3] H. Fernandez, S. Galar, M. Prati, R.C. Krawczyk, and F. Herrera (2018). *Learning from Imbalanced Data Sets.* Springer.

[4] J.H. Friedmann (2001). Greedy function approximation : a gradient boosting machine. *Ann. Stat., 29(5), 1189-1232.*

[5] T. Hastie, R. Tibshirani, and J. Friedman (2017). *The Elements of Statistical Learning.* Springer Series in Statistics.

[6] B.M. Hill (1975). A simple general approach to inference about the tail of a distribution. *Ann. Stat., v.3, 1163–1174.*

[7] A. Langousis, A. Mamalakis, M. Puliga, and R.) Deidda (2016). Threshold detection for the generalized pareto distribution : Review of representative methods and application to the noaa ncdc daily rainfall database. *AGU Publications.*

[8] S.M. Lundberg, G.G. Erion, and S.I. Lee (2019). Consistent individualized feature attribution for tree ensembles. *University of Washington.*

[9] A. Mornet (2021). Cours m2 actuariat : Théorie des valeurs extrêmes. ISFA.

[10] T.E. Raghunathan, J.M. Lepkowski, J. VanHoewyk, and P. Solenberger (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Statistique Canada, n12-001.*

Annexes

Annexe A

Différentes métriques

La norme euclidienne est définie de cette façon :

$$x \in \mathbb{R}^d, \|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}$$

La distance d'Hamming est définie de cette façon :

$$x, y \in \{0,1\}^d, d(x,y) = \sum_{i=1}^{d} \mathbb{1}_{x_i \neq y_i}$$

Annexe B

Théorème Centrale Limite

Soit $\{X_n\}_n$ une suite de variables aléatoires indépendantes et identiquement distribuées telle que $\mathbb{E}[X_1] = \mu < +\infty$ et $\mathbb{V}[X_1] = \sigma^2 \in \mathbb{R}_+^*$.
On a alors :

$$\mathbb{P}\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq x\right] \xrightarrow[n\to+\infty]{} \Phi(x)$$

avec Φ la fonction de répartition de la loi normale centrée réduite.

Annexe C

Algorithme des k-means

Soit un ensemble $\{X_1, ..., X_n\}$ de n individus. Soit $k \in \mathbf{N}^*$. L'algorithme de k-means consiste en le regroupement des n individus en k groupes (ou clusters) : $C_1, ..., C_k$ de manière à ce que les individus soient le plus proches possible au sein d'un cluster.
On cherche ainsi à minimiser la somme des distances au sein d'une classe :

$$\sum_{i=1}^{k} \sum_{X_j \in C_i} \|X_j - \mu_i\|^2$$

où $\mu_i = \dfrac{1}{card(C_i)} \sum\limits_{X_j \in C_i} X_j$ est le barycentre de C_i.

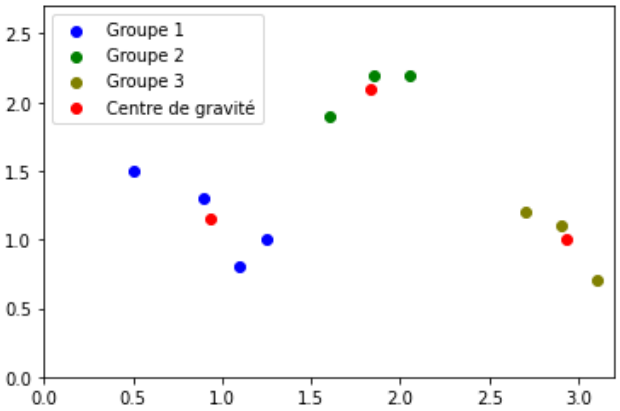

FIGURE C.1 – Illustration d'un clustering avec un 3-means

Annexe D

Rappels sur les arbres binaires

Un arbre binaire est une structure de données finie hiérarchisée dont les éléments sont appelés nœuds. Chaque nœud a au plus deux enfants qui sont eux aussi des nœuds appelés fils gauche et fils droit. Le nœud dont ils sont issus est quant à lui appelé nœud père. L'unique nœud n'ayant pas de père est appelé la racine et les nœuds n'ayant pas de fils sont appelés feuilles.

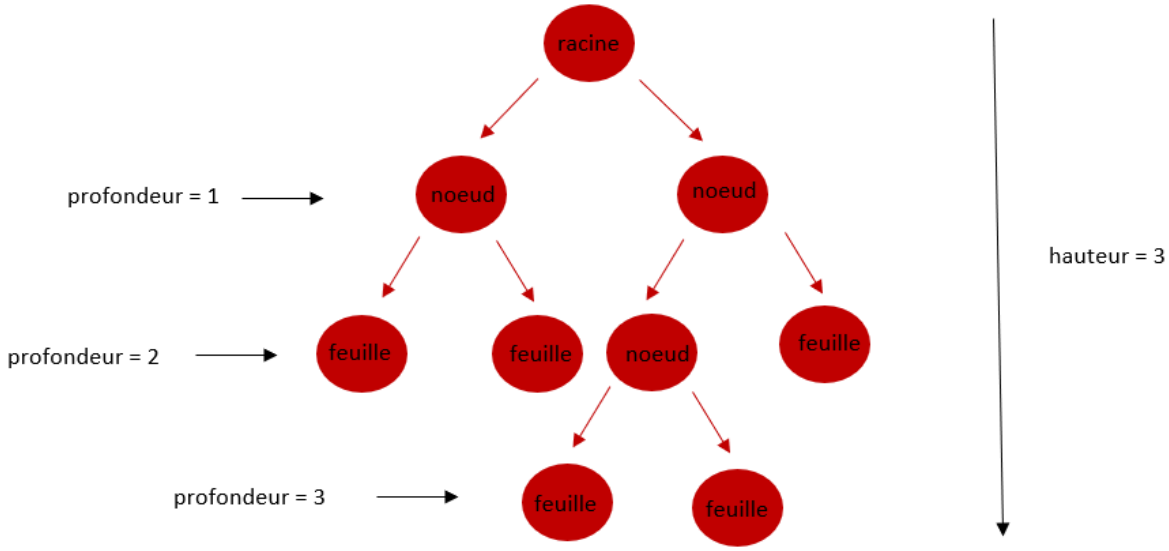

FIGURE D.1 – Illustration d'un arbre binaire

La profondeur d'un noeud est le nombre de liens qui le séparent de la racine. La hauteur de l'arbre est la profondeur maximale des feuilles.

Annexe E

La descente de gradient

On cherche à minimiser une fonction convexe $f : \mathbb{R} \longrightarrow \mathbb{R}$. On construit alors la suite x de cette manière :

$$x_0 \in \mathbb{R} \; ; \quad x_{k+1} = x_k - \lambda \cdot \nabla f(x_k)$$

jusqu'à ce que $|x_{k+1} - x_k| \leq \varepsilon$
L'idée est de se rapprocher du minimum en faisant des petits pas dans la bonne direction. La bonne direction étant ici le gradient. Le gradient indique la direction locale de croissance de la fonction. En prenant la direction opposée, on a plus de chance de se rapprocher du minimum.