Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires le 31 janvier 2024

Par : Benoît BERNARD
Titre : Étude du comportement de l'assuré au cours du processus de gestion et d'indemnisation d'un dossier de sinistre avec accent sur les sollicitations téléphoniques.

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : Covéa
Signature :

Directeur de Mémoire en entreprise :
Nom : Nicolas Brodeur
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

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

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature du candidat

Résumé

L'activité de gestion des sinistres et d'indemnisation est cruciale pour l'assureur. En effet, c'est lors de cette dernière que l'assureur peut évaluer précisément le coût des dégâts. Cependant, cette activité est elle-même soumise à un aléa qui impacte l'assureur à plusieurs titres. D'une part l'activité de gestion des sinistres et d'indemnisation participe aux coûts de gestion et peut donc influer sur le ratio combiné de l'assureur. D'autre part cette activité prend une part importante dans la relation client entre l'assureur et l'assuré en venant notamment impacter sa satisfaction ce qui a des répercussions sur sa fidélisation ainsi que sur l'image de marque de l'assureur. Une meilleure compréhension, maîtrise et modélisation de cette activité et des phénomènes qui la composent permettraient donc une meilleure maîtrise de l'activité d'assurance.

Au sein de l'activité de gestion et d'indemnisation les appels de gestion émis par l'assuré à destination de l'assureur sont particulièrement stratégiques. Ces derniers possèdent un coût non négligeable tout en étant source d'inconfort pour les deux parties car chaque appel fait intervenir un nouvel interlocuteur chez l'assureur. C'est pour ces raisons qu'il a été décidé de modéliser à l'échelle du dossier de sinistre la survenance de ces appels de gestion. Cela a notamment permis de montrer que la survenance d'un appel de gestion dans un futur proche est principalement dirigée par une activité récente sur le dossier.

Ce dernier résultat laisse à penser qu'une modélisation plus globale de l'activité de gestion et d'indemnisation devrait permettre d'en apprendre plus même sur le phénomène spécifique des appels de gestion. Cette approche bien que complexe à mettre en œuvre aura permis de dégager quelques résultats notamment sur le fait qu'une proactivité pourrait réduire la charge d'appels de gestion si elle est faite à l'écrit.

Les différents résultats obtenus au cours de cette étude devraient permettre une amélioration de l'activité de gestion et d'indemnisation. De plus, ce mémoire aura permis de mettre en évidence le potentiel qui existe à appliquer une approche actuarielle sur des pans de l'activité d'assurance pour lesquels cette approche est rarement utilisée.

Mots-clés : *indemnisation*, *comportement client*, *coûts de gestion*, *data science*, *processus de Hawkes*, *proactivité*.

Abstract

Claims management and indemnification are crucial activities for insurers. Indeed, it is during this activity that the insurer can accurately assess the cost of the damage. However, this activity is itself subject to a hazard that impacts the insurer in several ways. On the one hand, claims handling and compensation contribute to management costs, and can therefore influence the insurer's combined ratio. On the other hand, this activity plays an important role in the customer relationship between the insurer and the policyholder, notably by impacting customer satisfaction, which in turn has repercussions on customer loyalty and the insurer's brand image. A better understanding, control and modeling of this activity and the phenomena that make it up would therefore enable better control of the insurance business.

Within the management and claims activity, management calls made by the policyholder to the insurer are particularly strategic. They are not only costly, but also a source of discomfort for both parties, as each call involves a new interlocutor at the insurer. For these reasons, we decided to model the occurrence of these management calls at claims file level. In particular, this showed that the occurrence of a management call in the near future is mainly driven by recent activity on the file.

This last result suggests that a more global modeling of management and compensation activity should enable us to learn more even about the specific phenomenon of management calls. Although complex to implement, this approach has produced several results, notably that proactivity could reduce the management call load if it is done in writing.

The various results obtained during this study should lead to an improvement in management and claims activity. In addition, this dissertation has highlighted the potential of applying an actuarial approach to parts of the insurance business for which this approach is rarely used.

Keywords : *indemnification*, *customer behavior*, *management costs*, *data science*, *Hawkes process*, *proactivity*.

Note de Synthèse

Indemnisation et gestion des sinistres, les enjeux

Pour un assureur le moment de la gestion et de l'indemnisation des sinistres est crucial. C'est durant cette activité que l'assureur peut déterminer le coût des dégâts couverts par le contrat souscrit mais c'est également pendant cette période que sont engagés une part des coûts de gestion. Au delà des coûts le processus d'indemnisation et de gestion impacte l'expérience client. Si de nombreuses études actuarielles sont régulièrement menées sur le montant des dégâts des sinistres, il est assez rare qu'une approche statistique et mathématique soit adoptée pour traiter de l'indemnisation et de la gestion. Cette étude aura donc pour but de se focaliser sur ce processus d'indemnisation et de gestion en cherchant à le comprendre, le modéliser, l'analyser et le piloter dans le but d'éclairer l'assureur sur cette partie cruciale de son activité. Pour mener à bien cette étude on se concentrera sur un portefeuille de sinistres de fréquences de contrat multirisque habitation.

Les coûts de gestion sont parfois difficilement ventilable à l'échelle du dossier de sinistre mais lorsque cela est possible on se rend compte que ces derniers possèdent un coefficient de variation plus élevés que les coûts des dégâts. Cela signifie que ramener à une même échelle les coûts de gestion sont plus variables que ceux liés aux dégâts ce qui justifie d'autant plus d'adopter une analyse actuarielle sur le processus d'indemnisation et de gestion. L'évolution et la constitution de ce coût moyen par dossier est donnée dans la figure 1.

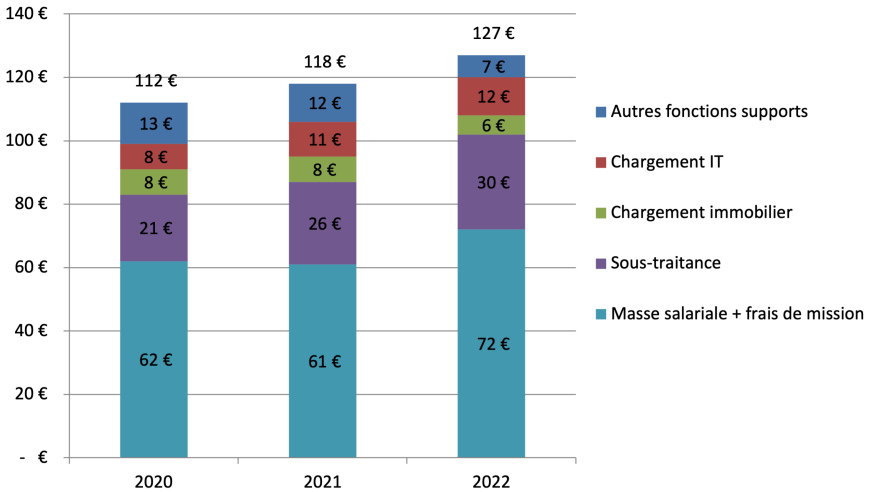


FIGURE 1 : Évolution et constitution du coût moyen de gestion

Il est possible d'y voir que le coût de gestion d'un sinistre connaît une tendance à la hausse, avec plusieurs facteurs clés contribuant à cette augmentation. Tout d'abord, la masse salariale représente

la part la plus importante de ces coûts mais aussi de leur augmentation. De plus, on observe une progression du poids de la sous-traitance dans les coûts de gestion. Quand on sait que les gestionnaires de sinistres passent une partie importante de leur journée de travail à traiter des appels de gestion on peut s'apercevoir que le traitement de ces appels participent grandement à la constitution des coûts de gestion, à hauteur de 15 € par appel. Si on ajoute à cela le fait ces appels ont un impact important sur la satisfaction client, il devient clair qu'ils doivent faire l'objet d'études plus approfondies. Sur le portefeuille de notre étude il a été estimé qu'entre 55% et 65% des appels de gestion réels étaient reliés à un dossier ce qui doit être pris en compte dans l'analyse des résultats.

Modélisation du phénomène d'appel de gestion par des outils de data science

Ayant été identifiés comme variable à fort enjeu on se propose dans un premier temps de modéliser l'occurrence future d'appels de gestion pour un dossier en cours de gestion. Plus précisément, soit t une date et d un dossier ouvert à la date t, soit AG_s^d le nombre total d'appels de gestion du dossier d intervenus avant s. On va chercher à modéliser la probabilité d'appel du dossier d en t, notée a_t^d, définit par

$$\begin{aligned} a_t^d :&= \mathbb{P}(AG_{t+14}^d > AG_t^d \mid \mathcal{F}_t^d) \\ &= \mathbb{E}(\mathbb{1}_{AG_{t+14}^d > AG_t^d} \mid \mathcal{F}_t^d), \end{aligned}$$

où $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ est une filtration qui contient l'évolution de la connaissances sur les informations (caractéristiques et événements) du dossier d. On est donc face à un problème de classification binaire et notre principale métrique de performance pour un modélisation sera l'AUC mais on utilisera également le nombre de dossiers appelants réellement parmi les 500 plus hautes prédictions (noté $DETEC_t(\cdot, 500)$) pour identifier la capacités des modèles à fournir un groupe de dossiers à haut risque si des mesures devaient être prises sur les dossiers aux plus hautes prédictions. Afin de réaliser des modélisations avec des outils classiques de data science il sera considéré que $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ regroupe les informations suivantes :

- la date d'ouverture du sinistre
- la typologie de sinistre
- le mode de gestion
- le mode de contact à la déclaration
- les dates et natures des 6 derniers événements connus (ouverture, appel de déclaration, de gestion ou sortant, pli reçu ou émis, ouverture d'un ordre de mission d'expertise ou de réparation)
- savoir si la prestation a été réalisée en t

La construction de la base de données des variables prédictives (représentant F_t^d) se fera notamment en représentant les dates des événements par la distance à la date de l'événement suivant ou par rapport à t pour le dernier événement connu.

En tant que premier modèle et modèle de référence on peut appliquer un régression logistique à notre problème de classification binaire qui suppose que pour une variable binaire à classifier Y en fonction de variables prédictives X on a

$$\mathbb{E}(Y|X) = \frac{1}{1+e^{-X^T\beta}}.$$

On évaluera également un arbre de classification ainsi que 3 autres modèles plus puissants mais moins explicables, à savoir une forêt aléatoire, un XGBoost et un réseau de neurones. Les performances de ces différents modèles sont données au tableau 1, il est possible d'y voir que l'ensemble des modèles ont des temps d'entraînement et de prédiction tout à fait compatible avec une utilisation récurrente et que les trois modèles moins explicables permettent d'améliorer légèrement les métriques évaluées sans que l'écart soit particulièrement important notamment avec la régression logistique.

modèle	entraînement	prédiction	AUC	$DETEC_{t_{test}}(\cdot, 500)$
régression logistique	51 secondes	$<$ 0,1 seconde	0,750	198
arbre de classification	39 secondes	$<$ 0,1 seconde	0,745	171
forêt aléatoire	3,08 minutes	2,1 secondes	0,761	205
XGBoost	4,03 minutes	$<$ 0,1 seconde	0,765	219
réseau de neurones	3,65 minutes	$<$ 0,1 seconde	0,764	217

TABLE 1 : Performances des modèles

Un fait marquant observé au cours de l'analyse de ces différents modèles est l'aspect auto-excitant des dossiers dans le sens ou la variable qui indique le temps écoulé depuis le dernier événement est systématiquement (et largement) la plus importante. Cela va dans le sens où plus un dossier a connu une activité récente plus sa probabilité d'être à l'origine d'un appel de gestion est élevée. Cette remarque laisse penser qu'il serait pertinent d'analyser le déroulement du phénomène de gestion dans sa globalité plutôt que de modéliser la survenance d'un événement donné.

Modélisation globale de la gestion d'un dossier

La dernière remarque du précédent chapitre conduit à s'intéresser à la dynamique globale de la gestion d'un dossier. Cette approche n'est pas incompatible avec la volonté de mieux comprendre le phénomène d'appel dans la mesure où une compréhension de l'occurrence de l'ensemble des événements ainsi que de leurs interactions devrait être en mesure d'apporter des précisions sur le phénomène particulier d'appel de gestion. L'idée est donc d'expliquer et de modéliser la survenance des différents événements d'un dossier au cours du temps. On pose le cadre suivant, soit d un dossier, on fixe sa date d'ouverture comme référence temporelle et on note E l'ensemble des événements possibles. On pose $\left(e_i^d, t_i^d\right)_{i\in\{1,\dots k_d\}}$ l'ensemble des couples $(\textit{événement}, \textit{date})$ du dossier d classé du plus ancien au plus récent avec k_d le nombre total d'événements du dossier d. Si on note T_d le temps que met le dossier d à être clôturé on a donc

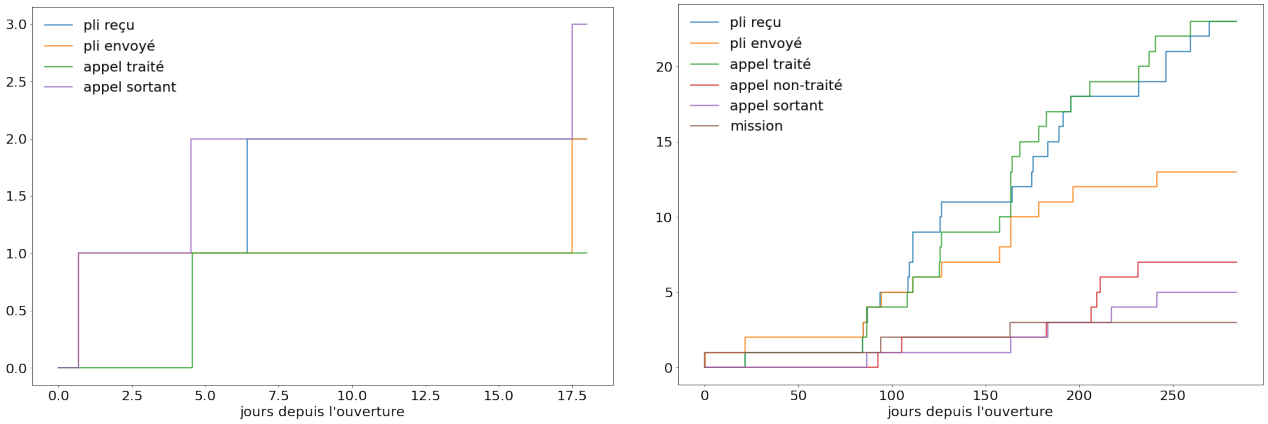
$$\left(e_1^d, t_1^d\right) = (\textit{ouverture}, 0) \text{ et } \left(e_{k_d}^d, t_{k_d}^d\right) = (\textit{clôture}, T_d)\,.$$

Soit $e \in E$, on définit à présent le processus de comptage suivant

$$N_t^{e,d} = \sum_{i=1}^{k_d} \mathbb{1}_{\left\{e_i^d = e,\ t_i^d \leq t\right\}} \ \forall t \geq 0. \tag{1}$$

La figure 2a donne un exemple de trajectoire des processus de chaque événement pour un dossier de notre étude ayant eu une gestion courte et relativement simple alors que la figure 2b représente un

dossier bien plus lourd en charge de gestion. Le processus de clôture n'a pas été représenté sur ces deux figures pour faciliter leur lecture, la date de clôture correspond à la dernière abscisse affichée.



(a) Un dossier simple

(b) Un dossier complexe

FIGURE 2 : Trajectoires de gestion

Ces deux exemples assez opposés laissent entrevoir des périodes d'accalmie communes à tous les processus ainsi que des concentrations de sauts sur des laps de temps très courts. Il est tout à fait logique d'observer ce deuxième phénomène car des plis peuvent être échangés au cours d'un appel ou encore car si un appel est interrompu ou non-traité il est probable qu'un autre le suive peu de temps après.

Une première approche pour modéliser les processus de comptage étudiés et de considérer qu'ils suivent un processus de Markov sur E. Cependant ce modèle semble peu pertinent pour représenter la trajectoire de gestion d'un dossier car un test de Kolmogorov-Smirnov au seuil de 5% permet de rejeter l'hypothèse de temps d'inter-arrivées exponentiels mais également car il impliquerait que la probabilité d'appel étudiée précédemment ne soit entièrement déterminée que par la nature du dernier événement hors un tel modèle ne permet d'obtenir un AUC que de 0,607.

Dans le but de mieux intégrer le phénomène d'auto-exciation précédemment observé on s'intéressera à une classe particulière de processus de comptage qui permettent de tenir compte des événements passés de façon que l'arrivée d'un événement viennent temporairement augmenter la probabilité d'un autre événement. On fait ici référence aux processus de Hawkes, qui trouvent de nombreuses applications pour la modélisation de phénomènes auto-excités dans des domaines variés comme l'épidémiologie, la finance, l'étude des répliques sismiques ou plus récemment en assurance pour modéliser le risque cyber ou terroriste. Les processus de Hawkes permettent de capter l'auto-excitation de processus de comptage au travers de la notion d'intensité d'un processus de comptage. Soit $(N_t)_{t>0}$ un processus de comptage tel que

$$N_t = \sum_{n\geq 0} \mathbb{1}_{T_n<t},$$

avec (T_n) les temps de saut du processus. On définit le processus d'intensité conditionnelle associé à $(N_t)_{t>0}$ noté λ par

$$\lambda(t) = \lim_{h\to 0} \mathbb{E}\left(\frac{N_{t+h} - N_t}{h} \middle| \mathcal{F}_t \right),$$

où la filtration $\mathcal{F}_t$ représente l'information connue sur le processus jusqu'au temps t exclu. L'intensité conditionnelle d'un processus représente donc la force avec laquelle le processus saute en t au sens où

pour dt petit le processus sautera sur $[t, t+\mathrm{d}t]$ avec probabilité $\lambda(t)\mathrm{d}t$. Dans le cas de la modélisation du processus de gestion l'intérêt d'une approche par processus de Hawkes serait donc de prendre en compte le fait que la survenance d'un événement augmente momentanément la probabilité d'une nouvelle survenance d'événement. Étant donné que le processus de gestion se compose de plusieurs processus de comptage on introduit les processus de Hawkes multivariés qu'on défini de la manière suivante. Soient $\left(N_t^1\right)_{t>0}, \ldots, \left(N_t^d\right)_{t>0}$ d processus de comptage (on note t_n^i les temps de sauts respectifs), on dit qu'ils forment un processus de Hawkes multivarié si les intensités conditionnelles adoptent la forme suivante

$$\lambda^i(t) = \mu^i + \sum_{j=1}^{d} \int_0^t \phi_{ij}(t-s)\mathrm{d}N_s^j = \mu^i + \sum_{j=1}^{d} \sum_{T_n^j < t} \phi_{ij}(t - T_n^j),$$

avec $\mu^i \geq 0$ les intensités de base et $\phi_{ij} : \mathbb{R}^+ \to \mathbb{R}^+$ les fonctions noyaux et on note Φ la matrice noyau qui en est composée. L'exemple de noyau le plus souvent utilisé de par sa simplicité et des avantages en termes de calibration des paramètres est le noyau exponentiel qui s'écrit sous la forme

$$\phi_{ij}(s) = \alpha_{ij}\beta_{ij}e^{-\beta_{ij}s},$$

avec $\alpha_{ij}, \beta_{ij} > 0, 1 \leq i, j \leq d$. Les paramètres peuvent s'interpréter de la manière suivante, α_{ij} indique avec quel force un saut du processus j vient perturber le processus i tandis que β_{ij} indique la manière dont cette perturbation va s'étaler dans le temps.

La figure 3 donne un exemple de simulation d'un processus de Hawkes en deux dimensions avec un noyau exponentiel tel que $\alpha_{11} = \alpha_{22} = 0.2$, $\alpha_{12} = \alpha_{21} = 0.4$, $\beta_{11} = \beta_{22} = \beta_{12} = \beta_{21} = 0.3$ et $\mu_1 = \mu_2 = 0.1$. On peut y observer les interactions entre les différents temps de sauts par le biais des intensités. On retrouve bien des périodes de calmes et de forte activité ce qui est une propriété souhaitée dans la modélisation du processus de gestion.

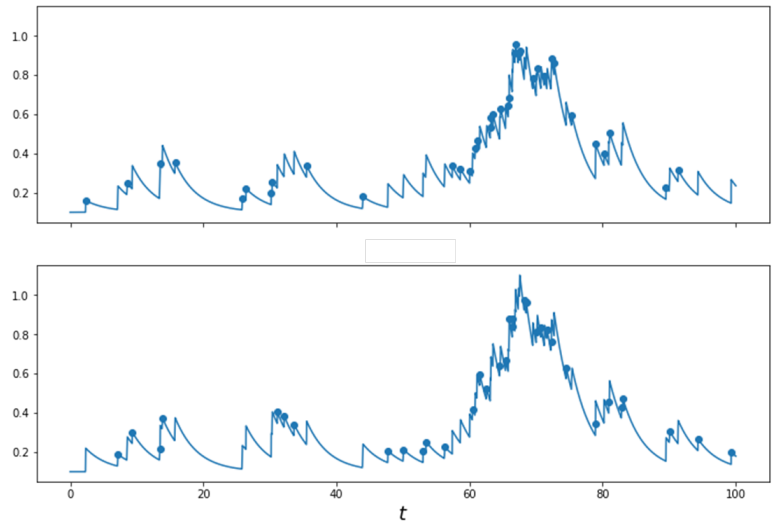


FIGURE 3 : Intensité et temps de sauts d'un processus de Hawkes à noyau exponentiel

L'application des processus de Hawkes à noyau exponentiel bien que semblant prometteuse n'aura permis de capter des interactions que de court terme (à l'échelle du jour). Le fait de considérer qu'après un laps de temps court les événements ne sont plus générés que par l'intensité de base est une limite dans notre modélisation car on se rapproche du cas markovien étudié précédemment. Le principe de

l'intensité conditionnelle d'un processus de Hawkes à noyau exponentiel a été réutilisé dans le cardre spécifique du phénomène d'appel de gestion en posant le modèle suivant,

$$a_t^d = \mathbb{P}(AG_{t+14}^d > AG_t^d \mid \mathcal{F}_t^d) = f\left(\mu + \sum_{j\in E}\sum_{T_n^j<t} \alpha_j\beta_j e^{-\beta_j(t-T_n^j)}\right),$$

avec E l'ensemble des événements, $\left(T_n^j\right)_{n\geq 0}$ les temps d'arrivées des événements de type j et f une fonction croissante à valeurs dans $[0,1]$. On peut considérer μ comme une intensité de base d'appel de gestion, α_j comme la force globale avec laquelle un événement de type j vient provoquer des appels de gestion et β_j comme caractérisant la décroissance de cet effet. L'utilité de la fonction f est de ramener l'intensité dans l'intervalle $[0,1]$ pour l'interpréter comme une probabilité. On peut choisir la fonction sigmoïde inverse du lien *logit*, soit $\sigma(x) := \frac{1}{1+e^{-x}}$, pour représenter l'intensité comme une probabilité.

Il est intéressant de constater que lorsqu'on fige les valeurs des β_j on se retrouve en présence d'un modèle de régression logistique. En effet, si on pose $X_j = \sum_{T_n^j<t} \beta_j e^{-\beta_j(t-T_n^j)}$ on peut alors réécrire l'équation comme

$$a_t = \sigma\left(\mu + \sum_{j\in E} \alpha_j X_j\right),$$

qui correspond bien à un modèle de régression logistique. Il est donc possible de voir la modélisation proposée comme une régression logistique possédant des hyper-paramètres $(\beta_j)_{\in E}$ qui viennent définir les variables prédictives utilisées. Ce constat va être utile pour calibrer les paramètres du modèle, on cherche les variables prédictives et donc les $(\beta_j)_{\in E}$ qui maximisent les performances de la régression logistique. On donne les paramètres du modèle dans le tableau 2. On peut constater que l'ouverture est l'événement qui excite le plus le processus de gestion ce qui n'est pas forcément surprenant quand on sait que les appels de gestions sont plus fréquents sur le début de la gestion d'un dossier et que cet événement ne peut se produire qu'une fois. Ensuite, on peut constater que ce sont les appels de gestion eux-mêmes qui se génèrent le plus ce qui va dans le sens des observations précédentes mais laisse toujours perplexes quant à l'utilité de ces appels répétés. Autre fait notable, l'envoi d'un pli ou l'ouverture d'un ordre de mission semble désexciter l'arrivée d'un appel de gestion. Une interprétation possible de ce dernier constat est que dans le cadre d'une mission le dossier va être entre les mains d'un autre professionnel (l'expert ou le réparateur) le temps de la mission et ne devrait donc pas donner lieu immédiatement à un contact avec le gestionnaire et que dans le cadre du pli envoyé on peut limiter les appels ayant uniquement pour but de connaître l'avancement du dossier. En ce qui concerne les appels sortants on pourrait s'attendre à observer un phénomène identique que pour les plis envoyés mais dans la pratique le client n'est pas forcément disponible au moment de l'appel et le simple fait d'avoir un appel manqué de son assureur peut le motiver à le rappeler. Sur l'ensemble de test on obtient une AUC de 0,728 et on détecte 170 appels à capacité donnée de 500. On peut considérer ce modèle comme un indicateur de la chaleur de l'activité d'un dossier.

	ouverture	appel d	appel g	appel s	mission	pli envoyé	pli reçu	μ
α_j	1.67	0.45	0.53	0.12	-0.23	-0.15	0.34	-3.15
β_j	0.12	0.02	0.02	0.05	0.15	0.07	0.10	

TABLE 2 : Estimation des paramètres du modèle d'intensité avec des β_j variables

Une dernière approche consistant à regrouper les événements d'un dossier en quelques séquences notamment en se basant sur des informations extraites des typologies (objets) des plis grâce à l'al-

gorithme de Word2vec a permis de poser un cadre général de modélisation qu'il pourrait être utile d'approfondir.

Test de priorisation

Dans le but de réduire la charge d'appels de gestion la possibilité d'identifier les dossiers à risque doit s'associer à des mesures permettant de les inhiber. La probabilité d'appel peut servir de score de priorisation tandis que les résultats du tableau 2 suggèrent que au-delà de la décision de créer un ordre de mission qui est un choix ayant un impact important sur le coût de gestion seul l'envoi d'un pli semble en mesure de freiner l'appel. Afin de mesurer l'impact de la prise de décision du gestionnaire on se propose de segmenter les dossiers à forte probabilité d'appel en fonction de l'action qui suit la date à laquelle a été évaluée cette probabilité. La segmentation se fera selon les possibilités suivantes.

- Aucune action n'est arrivée dans les deux semaines ou la prochaine action était un appel entrant.
- La première action dans les deux semaines était un pli envoyé.
- La première action dans les deux semaines était un appel sortant.
- La première action dans les deux semaines était un pli reçu.

Si on modélise la probabilité d'appel en rajoutant l'information de la nature de la prochaine action et qu'on identifie l'action minimisant la probabilité conditionnelle de chaque dossier on que presque systématiquement la meilleure prochaine action se joue entre attendre et envoyer un pli. Dans la pratique plus le dossier a une forte probabilité d'appel plus le modèle identifie l'envoi d'un pli comme étant la meilleure prochaine action ce qui peut se voir à la figure 4 où l'on représente la proportion de dossiers ayant pour meilleure prochaine action d'attendre ou d'envoyer un pli en fonction de la probabilité d'appel. Cette modélisation individuelle vient donc confirmer qu'une proactivité ayant pour but de limiter le nombre d'appels de gestion passe par la communication écrite. Un test de proactivité écrite a donc été initié en prenant les dossiers qui ont la plus forte probabilité d'appel à date et en les répartissant en trois groupes. Un groupe témoin servira uniquement pour l'analyse, un groupe fera l'objet d'une proactivité écrite avec traitement automatisé (un message générique) et un groupe fera l'objet d'une proactivité écrite avec traitement personnalisé.

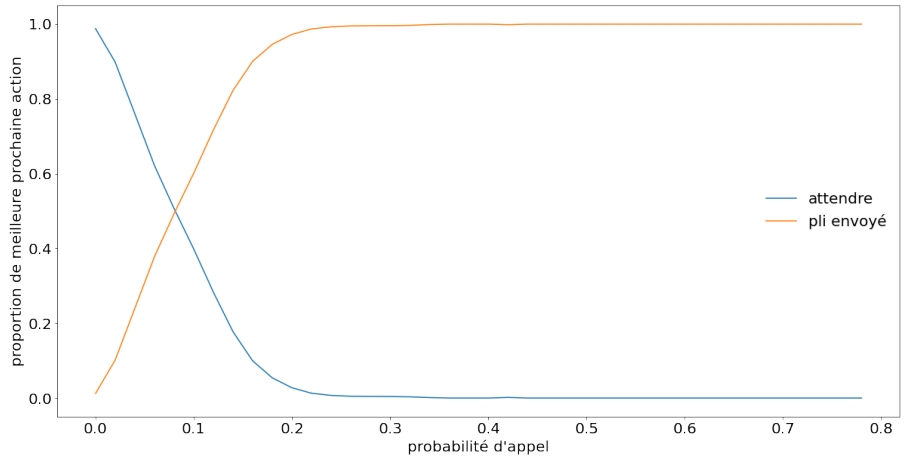


FIGURE 4 : Meilleure prochaine action en fonction de la probabilité d'appel

Synthesis note

Indemnification and claims management, the issues at stake

For an insurer, the claims management and settlement phase is crucial. It is during this activity that the insurer can determine the cost of the damage covered by the policy, but it is also during this period that a proportion of management costs are incurred. Beyond costs, the claims and management process impacts the customer experience. While many actuarial studies are regularly carried out on the amount of damage caused by claims, it is rather rare for a statistical and mathematical approach to be adopted to deal with compensation and management. The aim of this study is therefore to focus on the claims and management process, seeking to understand, model, analyze and manage it, in order to shed light on this crucial part of the insurer's business. To carry out this study, we will focus on a portfolio of frequency claims for multi-risk home insurance policies.

Management costs are sometimes difficult to break down on the scale of the claim file, but when this is possible, they have a higher coefficient of variation than damage costs. This means that, on the same scale, management costs are more variable than damage costs, all the more reason to adopt an actuarial analysis of the compensation and management process. The evolution and constitution of this average cost per case is shown in the figure 5.

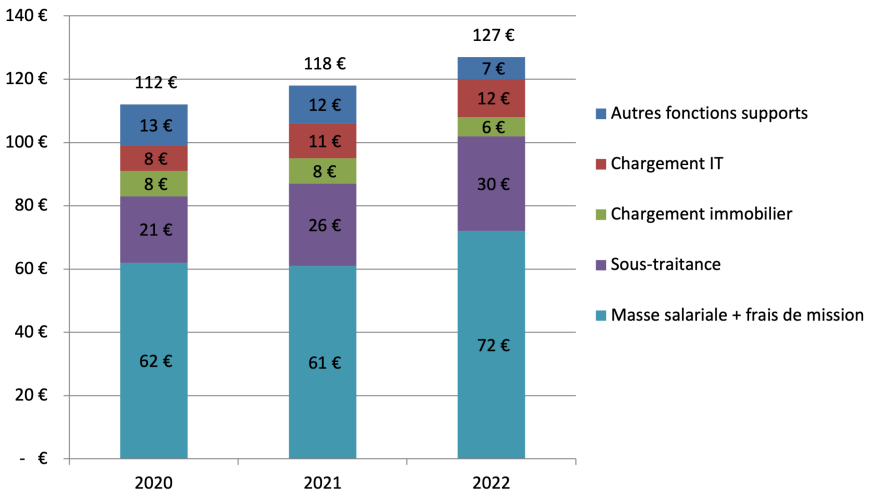


Figure 5: Evolution and constitution of average management costs

It can be seen that the cost of managing a claim is trending upwards, with several key factors contributing to this increase. Firstly, payroll accounts for the lion's share of these costs, and also of their increase. What's more, subcontracting is playing a growing role in management costs. When you consider that claims handlers spend a large part of their working day dealing with management

calls, you can see that handling these calls makes a major contribution to management costs, to the tune of 15 per call. If we add to this the fact that these calls have a major impact on customer satisfaction, it becomes clear that they need to be studied in greater depth. In our study portfolio, it was estimated that between 55% and 65% of actual management calls were case-related, which needs to be taken into account when analyzing results.

Modeling the management call phenomenon using data science tools

Having been identified as a high-stakes variable, we first propose to model the future occurrence of management calls for a file currently being managed. More precisely, let t be a date and d a file open on date t, and let AG_s^d be the total number of management calls for file d that occurred before s. We'll try to model the probability of calling file d in t, noted a_t^d, defined by

$$\begin{aligned} a_t^d &:= \mathbb{P}(AG_{t+14}^d > AG_t^d \mid \mathcal{F}_t^d) \\ &= \mathbb{E}(\mathbb{1}_{AG_{t+14}^d > AG_t^d} \mid \mathcal{F}_t^d), \end{aligned}$$

where $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ is a filtration containing the evolution of knowledge on the information (features and events) of the d folder. We are therefore faced with a binary classification problem, and our main performance metric for modeling will be the AUC, but we will also use the number of calling files actually among the 500 highest predictions (noted $DETEC_t(\cdot, 500)$) to identify the models' ability to provide a group of high-risk files if action were to be taken on the files with the highest predictions. In order to carry out modeling with classic data science tools, we will consider that $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ groups together the following information:

- the opening date of the claim
- type of claim
- management mode
- the declaration contact method
- the dates and nature of the last 6 known events (opening, declaration, management or outgoing call, letter received or issued, opening of an expert appraisal or repair order)
- find out if the service was performed in t.

The database of predictor variables (representing F_t^d) is built by representing event dates by the distance to the next event date, or relative to t for the last known event.

As a first and reference model we can apply logistic regression to our binary classification problem which assumes that for a binary variable to be classified Y as a function of predictor variables X we have

$$\mathbb{E}(Y|X) = \frac{1}{1+e^{-X^T\beta}}.$$

We'll also evaluate a classification tree and 3 other more powerful but less explainable models, namely a random forest, an XGBoost and a neural network. The performance of these different models is shown in table 3, where we can see that all the models have training and prediction times that are perfectly compatible with recurrent use, and that the three less-explainable models enable us to slightly improve the metrics evaluated, without the gap being particularly large, notably with logistic regression.

model	training	prediction	AUC	$DETEC_{t_{test}}(\cdot, 500)$
logistic regression	51 seconds	$<$ 0,1 second	0,750	198
classification tree	39 seconds	$<$ 0,1 second	0,745	171
random forest	3,08 minutes	2,1 seconds	0,761	205
XGBoost	4,03 minutes	$<$ 0,1 second	0,765	219
neural network	3,65 minutes	$<$ 0,1 second	0,764	217

Table 3: Model performance

A striking fact observed during the analysis of these different models is the self-exciting aspect of the files, in the sense that the variable indicating the time elapsed since the last event is systematically (and largely) the most important. This suggests that the more recent the activity of a file, the higher its probability of being at the origin of a management call. This suggests that it would be more appropriate to analyze the management phenomenon as a whole, rather than modeling the occurrence of a given event.

Global modeling of file management

The last remark of the previous chapter leads us to look at the overall dynamics of case management. This approach is not incompatible with the desire to better understand the call phenomenon, insofar as an understanding of the occurrence of all events and their interactions should be able to shed light on the specific phenomenon of management calls. The idea is therefore to explain and model the occurrence of different events in a file over time. Let d be a file, with its opening date as a time reference, and let E be the set of possible events. Let $\left(e_i^d, t_i^d\right)_{i\in\{1,\dots k_d\}}$ be the set of $(event, date)$ pairs in folder d, ranked from oldest to most recent, with k_d the total number of events in folder d. If we note T_d the time it takes for file d to be closed, then we have

$$\left(e_1^d, t_1^d\right) = (opening, 0) \text{ and } \left(e_{k_d}^d, t_{k_d}^d\right) = (closing, T_d)\,.$$

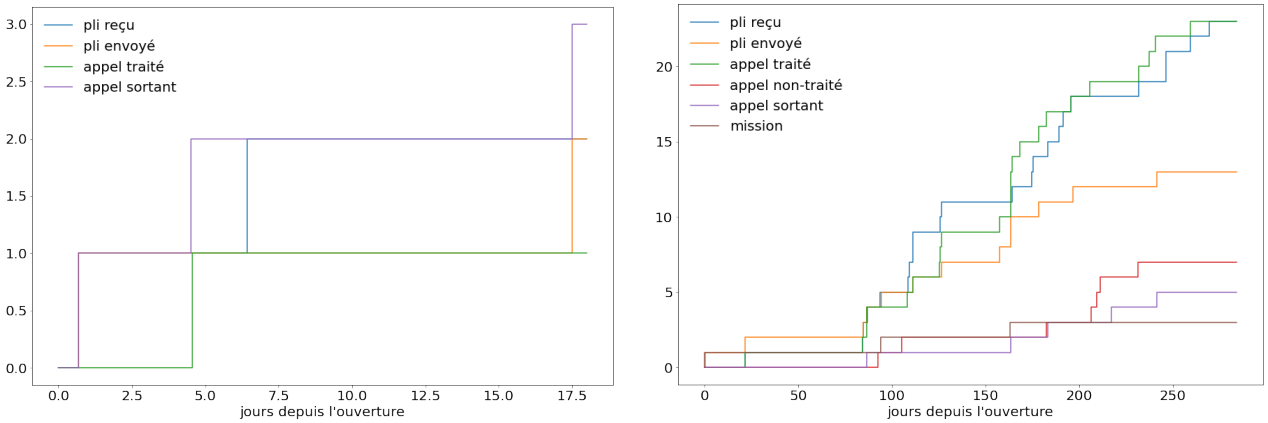
Let $e \in E$, we now define the following counting process

$$N_t^{e,d} = \sum_{i=1}^{k_d} \mathbb{1}_{\left\{e_i^d = e,\ t_i^d \leq t\right\}} \ \forall t \geq 0. \tag{2}$$

Figure 6a gives an example of the process trajectory for each event for a file in our study with a short and relatively simple management cycle, while figure 6b represents a file with a much heavier management load. The closing process has not been represented on these two figures to make them easier to read; the closing date corresponds to the last abscissa displayed.

These two rather opposite examples suggest periods of lull common to all processes, as well as concentrations of jumps over very short periods of time. It's quite logical to observe the latter phenomenon, since folds can be exchanged during a call, or if one call is interrupted or not processed, another is likely to follow shortly afterwards.

A first approach to modeling the counting processes studied is to consider that they follow a Markov process on E. However, this model does not seem very appropriate for representing the management trajectory of a file, as a Kolmogorov-Smirnov test with a threshold of 5% rejects the hypothesis of exponential inter-arrival times, and also because it would imply that the call probability studied previously is entirely determined only by the nature of the last event, out of which such a model achieves an AUC of only 0.607.



(a) A simple file

(b) A complex file

Figure 6: Management trajectories

With the aim of better integrating the self-excitation phenomenon observed above, we'll be looking at a particular class of counting processes that allow past events to be taken into account, so that the arrival of one event temporarily increases the probability of another. We are referring here to Hawkes processes, which have numerous applications for modeling self-exciting phenomena in fields as varied as epidemiology, finance, the study of seismic aftershocks or, more recently, in insurance to model cyber or terrorist risk. Hawkes processes capture the self-excitation of counting processes through the notion of the intensity of a counting process. Let $(N_t)_{t>0}$ be a counting process such that

$$N_t = \sum_{n \geq 0} \mathbb{1}_{T_n < t},$$

with (T_n) the process jump times. We define the conditional intensity process associated with $(N_t)_{t>0}$ denoted λ by

$$\lambda(t) = \lim_{h \to 0} \mathbb{E}\left(\frac{N_{t+h} - N_t}{h} \middle| \mathcal{F}_t \right),$$

where the filtration $\mathcal{F}_t$ represents the information known about the process up to time t excluded. The conditional intensity of a process thus represents the strength with which the process jumps in t in the sense that for dt small the process will jump on $[t, t+textdt]$ with probability $ambda(t)$dt. In the case of modeling the management process, the interest of a Hawkes process approach would therefore be to take into account the fact that the occurrence of an event momentarily increases the probability of a new event occurring. Given that the management process is made up of several counting processes, we introduce multivariate Hawkes processes, defined as follows. If $\left(N_t^1\right)_{t>0}, \ldots, \left(N_t^d\right)_{t>0}$ d counting processes (note t_n^i the respective jump times), they are said to form a multivariate Hawkes process if the conditional intensities adopt the following form

$$\lambda^i(t) = \mu^i + \sum_{j=1}^{d} \int_0^t \phi_{ij}(t-s) \mathrm{d}N_s^j = \mu^i + \sum_{j=1}^{d} \sum_{T_n^j < t} \phi_{ij}(t - T_n^j),$$

with $\mu^i \geq 0$ the basic intensities and $\phi_{ij} : \mathbb{R}^+ \to \mathbb{R}^+$ are the kernel functions, and Φ is the kernel matrix composed of them. The most commonly used kernel, due to its simplicity and advantages in terms of parameter calibration, is the exponential kernel, written as

$$\phi_{ij}(s) = \alpha_{ij} \beta_{ij} e^{-\beta_{ij} s},$$

with $\alpha_{ij}, \beta_{ij} > 0, 1 \leq i, j \leq d$. The parameters can be interpreted as follows: α_{ij} indicates how strongly a jump in process j disturbs process i, while β_{ij} indicates how this disturbance is spread out over time.

Figure 7 gives an example of a simulation of a Hawkes process in two dimensions with an exponential kernel such that $\alpha_{11} = \alpha_{22} = 0.2$, $\alpha_{12} = \alpha_{21} = 0.4$, $\beta_{11} = \beta_{22} = \beta_{12} = \beta_{21} = 0.3$ and $\mu_1 = \mu_2 = 0.1$. The interactions between the different jump times can be seen through the intensities. There are periods of calm and periods of high activity, which is a desirable property when modeling the management process.

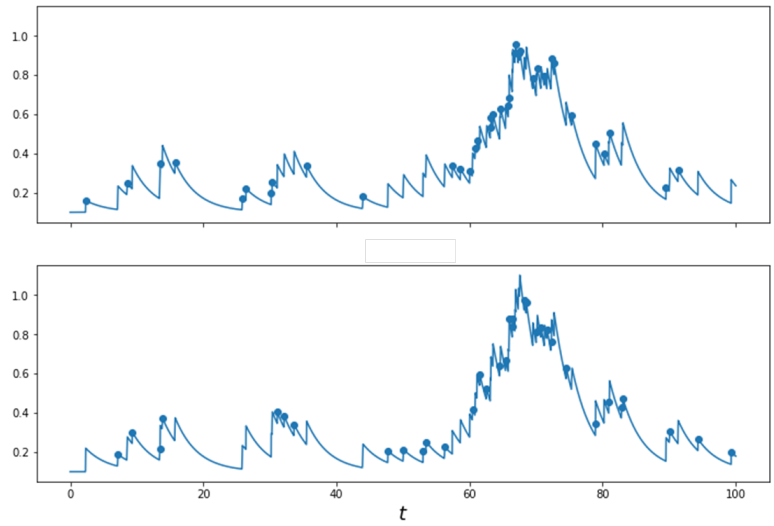


Figure 7: Intensity and jump times of a Hawkes process with exponential kernel

The application of Hawkes processes with exponential kernels, although promising, will only capture short-term interactions (on a daily scale). Considering that after a short period of time, events are only generated by the basic intensity is a limitation in our modeling, as we are approaching the Markovian case studied previously. The principle of the conditional intensity of a Hawkes process with an exponential kernel has been reused in the specific carder of the management call phenomenon by positing the following model,

$$a_t^d = \mathbb{P}(AG_{t+14}^d > AG_t^d \mid \mathcal{F}_t^d) = f\left(\mu + \sum_{j \in E} \sum_{T_n^j < t} \alpha_j \beta_j e^{-\beta_j (t - T_n^j)}\right),$$

with E the set of events, $\left(T_n^j\right)_{n \geq 0}$ the arrival times of events of type j and f an increasing function with values in $[0, 1]$. We can think of μ as a basic intensity of management calls, α_j as the overall strength with which an event of type j provokes management calls, and β_j as characterizing the decay of this effect. The purpose of the f function is to bring the intensity into the $[0, 1]$ interval, so that it can be interpreted as a probability. We can choose the inverse sigmoid function of the *logit* link, i.e. $\sigma(x) := \frac{1}{1+e^{-x}}$, to represent intensity as a probability.

Interestingly, when we freeze the β_j values, we find ourselves in the presence of a logistic regression model. Indeed, if we pose $X_j = \sum_{T_n^j < t} \beta_j e^{-\beta_j (t - T_n^j)}$ we can then rewrite the equation as

$$a_t = \sigma\left(\mu + \sum_{j \in E} \alpha_j X_j\right),$$

which corresponds to a logistic regression model. It is therefore possible to see the proposed model as a logistic regression with hyper-parameters $(\beta_j)_{\in E}$ that define the predictor variables used. This observation will be useful for calibrating the model parameters: we're looking for the predictor variables and therefore the $(\beta_j)_{\in E}$ that maximize the performance of the logistic regression. The model parameters are given in table 4. We can see that opening is the event that most excites the management process, which isn't necessarily surprising given that management calls are most frequent at the start of a file's management and that this event can only occur once. Secondly, we can see that it's the management calls themselves that are generated the most, which is in line with the previous observations, but still leaves us wondering about the usefulness of these repeated calls. Another noteworthy fact is that sending a letter or opening a mission order seems to de-energize the arrival of a management call. One possible interpretation of this last observation is that, in the context of an assignment, the file will be in the hands of another professional (the expert or the repairer) for the duration of the assignment and should therefore not immediately give rise to contact with the manager, and that, in the context of the envelope sent, we can limit calls whose sole purpose is to find out about the progress of the file. As for outgoing calls, we might expect to observe the same phenomenon as for outgoing mail, but in practice the customer is not necessarily available at the time of the call, and the simple fact of having a missed call from his insurer may motivate him to call back. On the test set, we obtained an AUC of 0.728 and detected 170 calls at a given capacity of 500. We can consider this model as an indicator of the heat of a file's activity.

	opening	call d	call m	call out	mission	letter sent	letter received	μ
α_j	1.67	0.45	0.53	0.12	-0.23	-0.15	0.34	-3.15
β_j	0.12	0.02	0.02	0.05	0.15	0.07	0.10	

Table 4: Estimation of intensity model parameters with variable β_j.

A final approach, which consists of grouping the events in a file into a few sequences based on information extracted from the folds' typologies (objects) using the Word2vec algorithm, has provided a general modeling framework that could usefully be developed further.

Prioritization test

In order to reduce the burden of management calls, the ability to identify high-risk files must be combined with measures to inhibit them. The probability of a call can be used as a prioritization score, while the results of the table 4 suggest that beyond the decision to create a mission order, which is a choice with a significant impact on management costs, only the sending of a letter seems capable of curbing the call. In order to measure the impact of the manager's decision, we propose to segment files with a high probability of appeal according to the action that follows the date on which this probability was assessed. The segmentation will be based on the following possibilities.

- No action occurred within two weeks, or the next action was an incoming call.
- The first action within two weeks was to send a letter.
- The first action within two weeks was an outgoing call.
- The first action within two weeks was an incoming letter.

If we model the call probability by adding information about the nature of the next action, and identify the action that minimizes the conditional probability of each file, we can see that almost systematically the best next action is between waiting and sending a letter. In practice, the higher the probability of appeal, the more the model identifies sending an envelope as the best next action, as shown in figure 8, where we plot the proportion of cases whose best next action is to wait or to send an envelope, as a function of the probability of appeal. This individual modeling confirms that proactivity aimed at limiting the number of management calls requires written communication. A test of written proactivity was therefore initiated by taking the files with the highest probability of call to date and dividing them into three groups. A control group will be used for analysis purposes only, a group will be subjected to written proactivity with automated processing (a generic message) and a group will be subjected to written proactivity with personalized processing.

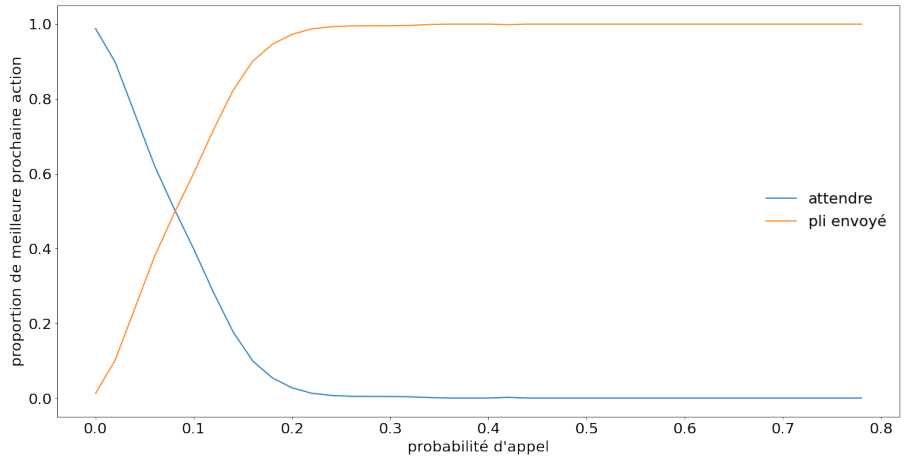


Figure 8: Best next action based on call probability

Remerciements

Je tiens à remercier sincèrement tous ceux qui ont pu m'aider à réaliser ce mémoire. Tout particulièrement j'adresse mes remerciements à Nicolas Brodeur, mon directeur de stage, pour le temps et l'expertise qu'il a su m'accorder et à Christophe Dutang, mon tuteur académique, pour ses conseils et relectures. Je tiens également à remercier Maxence Jeunesse qui malgré un emploi du temps chargé a trouvé le temps de m'accompagner et m'apporter de précieuses idées.

Table des matières

Introduction

L'indemnisation est un moment crucial dans l'activité d'assurance. En effet, c'est lors de cette étape que sont engagés une part des coûts de gestion mais aussi que se joue la satisfaction de l'assuré et donc sa fidélisation. Les échanges et activités ayant lieu entre la déclaration par l'assuré et la clôture d'un dossier sinistre sont donc des moments clés en s'inscrivant dans cette double problématique de coûts de gestion et de satisfaction client. Il est à noter que la survenance de ces échanges au cours du processus de gestion et d'indemnisation revêt un caractère aléatoire tant à l'échelle du dossier de sinistre que d'un point de vue agrégé. Bien qu'étant de nature aléatoire tout en impactant son activité et ses coûts, ce processus de gestion et d'indemnisation ainsi que les échanges qu'il comporte restent peu étudiés sous une approche actuarielle par les assureurs.

Au sein de ce processus de gestion et d'indemnisation les appels de gestion qui sont émis par l'assuré à destination de l'assureur après que la déclaration a été faite sont un enjeu d'envergure. En effet, pour être traités ces derniers demandent qu'un gestionnaire soit instantanément disponible tout en rendant impossible le suivi de l'activité d'un dossier par un même gestionnaire de son ouverture jusqu'à sa clôture. Au delà de cet inconfort généré tant du côté assureur que du côté assuré, les capacités mises en place pour faire face à ces sollicitations téléphoniques engendrent des coûts importants notamment par la mise en place de sous-traitance visant à absorber les variations d'activité.

Que ce soit pour obtenir une meilleure maîtrise de son activité et de ses coûts de gestion mais aussi pour renforcer la satisfaction client l'assureur a intérêt à étudier, modéliser, comprendre, prédire et piloter son processus de gestion et d'indemnisation de la même façon qu'il pourrait le faire pour la sinistralité. Ce besoin spécifique, porté par la MAAF et orienté sur les appels de gestion, sera étudié au cours de ce mémoire.

L'opportunité de diminuer ses coûts de gestion tout en améliorant la satisfaction client devrait à terme permettre à l'assureur de proposer une offre plus attractive sur un marché où la concurrence impose d'exploiter toutes les opportunités d'amélioration de l'offre que ce soit en termes de tarifs ou de qualité de service.

Ce mémoire entend appliquer une analyse actuarielle au processus de gestion et d'indemnisation en s'attachant notamment à la modélisation des phénomènes aléatoires, et donc à risque, qui le compose. Dans la mesure où les enseignements de ce mémoire ont pour vocation à amener des changements dans la gestion des dossiers de sinistres on travaillera à l'échelle du dossier en s'intéressant à la survenance des événements, et en particulier les appels de gestion, au sein d'un dossier. On ne travaillera donc pas sur la survenance des sinistres mais sur leur gestion qui peut être vue comme une sous sinistralité qui reste aléatoire après la déclaration.

Au cours de ce mémoire nous commencerons par un chapitre introductif 1 rappelant rapidement en quoi consiste le processus d'indemnisation ses enjeux, l'importance des flux téléphoniques ainsi que le

cadre de l'étude et les données utilisées. Le second chapitre 2 sera dédié à l'utilisation de certains outils de data science pour modéliser l'imminence d'un appel de gestion à l'échelle du dossier. Le troisième chapitre 3 de ce mémoire sera consacré à une approche plus globale sur la modélisation de l'ensemble de la gestion bien que cette tâche soit plus difficile à mettre en œuvre. Enfin, un dernier chapitre 4 viendra présenter quelques utilisations possibles des résultats obtenus tout au long de l'étude avec notamment la mise en place d'un test de proactivité visant à réduire la charge de gestion.

L'ensemble des résultats obtenus dans ce mémoire sont à visée exploratoire mais permettent de mettre en avant les opportunités qu'apporte la mise en place d'une approche actuarielle dans d'autres domaines de l'assurance que ceux habituellement étudiés.

Chapitre 1

Le processus d'indemnisation, contexte, pratiques et enjeux des échanges téléphoniques

Dans ce premier chapitre il sera abordé le fonctionnement du processus de gestion de sinistre et d'indemnisation au sein de la branche Incendie Risques Divers (IRD) des particuliers (PRI) au sein de la Direction Indemnisation de la MAAF ainsi que l'envergure des coûts de gestion. Suite à cela une attention particulière sera portée sur l'importance des échanges téléphoniques. Enfin, le cadre de l'étude, les données à disposition ainsi que les objectifs de modélisation seront explicités.

1.1 Contexte et description du processus d'indemnisation et de gestion de sinistre

1.1.1 Indemnisation et gestion, généralités

Cette section est inspirée de Brodeur (2019). Le métier d'assureur repose essentiellement sur la notion d'indemnisation des sinistres subis par ses assurés. Le caractère qui distingue ce métier des autres activités commerciales est son cycle de production inversé. Concrètement, une société d'assurance collecte les primes de ses clients en amont, afin de constituer des provisions destinées à couvrir les sinistres qui surviendront ultérieurement.

L'étape de l'indemnisation revêt une importance cruciale à la fois pour la maîtrise des coûts de l'entreprise et pour la relation avec le client. En effet, c'est à ce moment précis que le client évalue si les engagements et les promesses formulés par l'assureur lors de la souscription du contrat sont réellement tenus.

Pour l'assureur, il s'agit alors de trouver un équilibre délicat entre deux objectifs souvent opposés. D'un côté, il cherche à maîtriser ses coûts en proposant des indemnisations qui peuvent sembler "peu généreuses". Cette approche vise à assurer une rentabilité à court terme pour l'entreprise, en limitant les dépenses liées aux sinistres. Cependant, cette approche comporte le risque de susciter la méfiance ou le mécontentement chez les assurés, ce qui pourrait nuire à la relation commerciale à long terme.

D'un autre côté, l'assureur doit également aspirer à offrir un service client d'excellence. En répondant de manière satisfaisante et équitable aux attentes des assurés lors de l'indemnisation, il renforce la confiance et la fidélité de ces derniers envers la marque. Cette stratégie favorise une rentabilité à

long terme, en consolidant les liens commerciaux et en créant une image positive de la compagnie d'assurance.

Ainsi, l'assureur est confronté à un dilemme constant : d'une part, il doit gérer les coûts de manière rigoureuse pour assurer sa viabilité financière, et d'autre part, il doit veiller à satisfaire pleinement les attentes de ses clients afin de préserver une relation client solide et durable. Trouver le juste équilibre entre ces deux impératifs est une tâche complexe qui nécessite une expertise et une gestion habile de la part de la société d'assurance.

Le processus d'indemnisation et de gestion des sinistres reste une thématique peu étudiée sous un angle actuariel et comme une variable risquée. Cependant, loin d'être uniformes les coûts de gestion peuvent représenter une part non négligeable du coût total d'un sinistre, à titre d'exemple pour un sinistre rentrant dans le cadre de cette étude, les sinistres IRD PRI avec des dégâts inférieurs à 23 000 € (voir justification plus loin), le coût moyen de gestion était de 127 € en 2022. Sur des sinistres nombreux et peu coûteux ces coûts de gestions viennent participer de manière significative à l'aléa du ratio combiné. Le processus de gestion et d'indemnisation d'un sinistre en assurance peut varier d'une compagnie d'assurance à une autre, mais voici une vue d'ensemble générale du processus tel qu'il est pratiqué par la MAAF sur les sinistres IRD PRI :

- Déclaration du sinistre : L'assuré doit informer son assureur de l'incident ou du sinistre dès que possible. Cela peut être fait par téléphone, en ligne, par écrit ou lors d'une visite en agence.
- Mise en jeu des garanties : Le gestionnaire chargé du dossier doit examiner attentivement le contrat d'assurance souscrit par le client. Ce contrat détaille les différentes garanties auxquelles le client a droit en fonction de ses besoins spécifiques et des options choisies. Le gestionnaire vérifie donc d'abord si le sinistre en question est couvert par les garanties incluses dans le contrat. Si le sinistre n'est pas couvert ou s'il fait l'objet d'une exclusion spécifique, l'assureur ne prendra pas en charge les dommages déclarés, et le client ne sera pas indemnisé. On parle alors de "sinistre sans suite totale", car aucune compensation ne sera versée. Dans près de la moitié des cas (sur notre périmètre d'étude) le dossier sera classé sans suite.

  Il est à noter que ces dossiers sans suite ont un coût de gestion que l'assureur doit assumer même si le sinistre n'est pas couvert. Ce coût de gestion peut devenir important si les échanges s'étalent au-delà de la déclaration (ce qui est le cas dans 70% des dossiers) et surtout dans les cas où un expert a été missionné. En revanche, si le gestionnaire estime que le sinistre est couvert par le contrat d'assurance, il poursuit le traitement du dossier. Il recueille toutes les informations pertinentes concernant les dommages déclarés par le client, comme la nature des dommages, leur étendue et les circonstances de l'incident.
- Évaluation du sinistre : Le gestionnaire du sinistre évalue les dommages ou les pertes subis par l'assuré. Cela peut impliquer une visite sur les lieux par des experts salariés ou libéraux, l'obtention de témoignages ou de documents pertinents, et la collecte de preuves pour déterminer l'étendue des dommages. Pour établir une première estimation approximative du montant d'indemnisation à prévoir, le gestionnaire utilise un outil d'aide à l'évaluation (OAE). Cet outil est conçu pour faciliter le processus d'évaluation des dommages et fournir une première estimation rapide et grossière du coût probable de l'indemnisation. Cependant, il est essentiel de noter que l'estimation fournie par l'OAE n'est qu'une évaluation préliminaire et approximative. Le montant final d'indemnisation peut être ajusté notamment après une évaluation plus détaillée des dommages par des experts.
- Proposition d'indemnisation : Dans le cas où il n'y a pas d'expertise l'assureur propose à l'assuré un mode d'indemnisation qui peut être de plusieurs natures.

  - L'indemnisation gré à gré, qui correspond à un versement monétaire clôturant le dossier, est surtout utilisée pour les sinistres dont les montants s'élèvent à quelques centaines d'euros car elle permet de contenir le coût du sinistre étant donné que des propositions inférieures à ce que coûterait une réparation par un professionnel sont souvent avantageuses pour l'assuré. Ce mode d'indemnisation est également plébiscité car il permettrait de réduire les coûts de gestion.
  - L'indemnisation sur devis-facture reste la plus répandue. Elle correspond au cas où l'assuré se charge de faire réparer les dégâts du sinistre par un professionnel de son choix après acceptation du devis par l'assureur qui réglera la facture après plusieurs acomptes éventuels.
  - La réparation en nature est un mode d'indemnisation qui n'occasionne pas de versement monétaire de l'assureur vers l'assuré. Au lieu de cela l'assureur propose de faire intervenir un réparateur en nature partenaire de l'entreprise qui prendra rendez-vous avec l'assuré afin de réaliser les travaux de réparation et/ou de remplacement. L'assureur réglera directement le réparateur mais engage sa responsabilité sur la réalisation et la qualité de l'intervention ce qui n'est pas le cas lors d'une indemnisation en devis-facture. En ayant des partenariats avec les réparateurs, l'assureur peut se permettre de négocier des tarifs préférentiels.

- L'assureur procède à la clôture du dossier lorsque la proposition d'indemnisation est acceptée et réalisée et que les éventuels recours et traitements administratifs ont été réalisés.

Il est important de souligner que l'ensemble des étapes ci-dessus peuvent être réalisées plusieurs fois et dans un ordre varié. Par exemple, un sinistre peut être classé sans suite après le passage d'un expert (on parle de "sans suite indemnitaire"). C'est cette succession d'étapes propre à chaque dossier qui constitue le processus de gestion et d'indemnisation. Ce processus donne lieu à de nombreux échanges entre l'assuré et le gestionnaire qui constituent l'essentiel de l'activité d'un gestionnaire et donc a fortiori du coût de gestion.

1.1.2 Enjeux et risques du coût de gestion

Si de nombreuses études actuarielles portent sur l'évaluation du risque que constitue le montant des dégâts d'un sinistre en assurance il n'en existe que peu qui font cette démarche sur les coûts de gestion. Ce constat peut surprendre dans la mesure où ces coûts de gestion tout comme ceux relatifs aux dégâts du sinistre participent aux charges de l'assureur et doivent être prise en compte pour la tarification et le provisionnement. Si ces frais de gestion sont parfois assimilés à une constante ou considérés comme proportionnels au coût du sinistre car d'apparence de nature moins risquée il peut également être fastidieux voire impossible d'aller les déterminer à l'échelle du dossier de sinistre. Cependant, pour illustrer la nécessité de changer ces pratiques, en se basant sur les sinistres IRD PRI, une comparaison des coûts de gestion et des dégâts va être effectuée. Afin de comparer le risque de deux variables aléatoires nous allons utiliser la notion de coefficient de variation.

Définition 1.1 (Coefficient de variation)**.** Le coefficient de variation (CV) d'une variable aléatoire X est défini comme le ratio de son écart-type σ_X sur sa moyenne μ_X, $CV = \frac{\sigma_X}{\mu_X}$.

Les coûts des dégâts Sur la base d'un portefeuille de dossiers de sinistres ouverts en 2021 et 2022 dans le cadre IRD PRI 45,7% des dossiers n'ont pas engendrés de coût liés aux dégâts (sans suite et sans suite indemnitaire). Pour les 54,3% des dossiers ayant donné lieu à un versement ou une prestation de l'assureur à destination de l'assuré la distribution des coûts est donnée dans la figure 1.1.

La moyenne de ces coûts liés aux dégâts (y compris sans suite et sans suite indemnitaire) est de $\mu_{dégâts} = 536$ € et le coefficient de variation de $CV_{dégâts} = 2,37$.

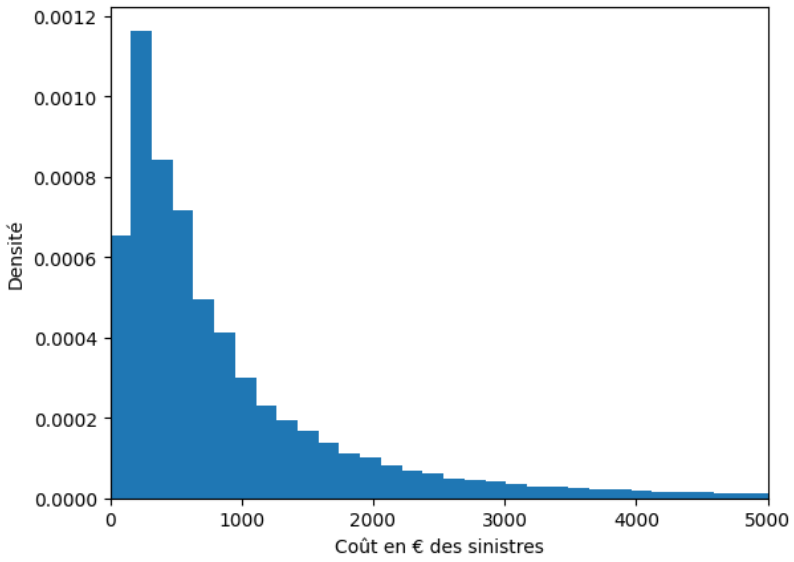


FIGURE 1.1 : Distribution des coûts liés aux dégâts

Les coûts de gestion Comme évoqué ci-dessus les coûts de gestion par dossier peuvent être durs à identifier. Néanmoins, il est possible d'estimer un coût de gestion moyen via des données agrégées qui sont difficilement ventilables par dossier. L'évolution et la constitution de ce coût moyen par dossier est donnée dans la figure 1.2.

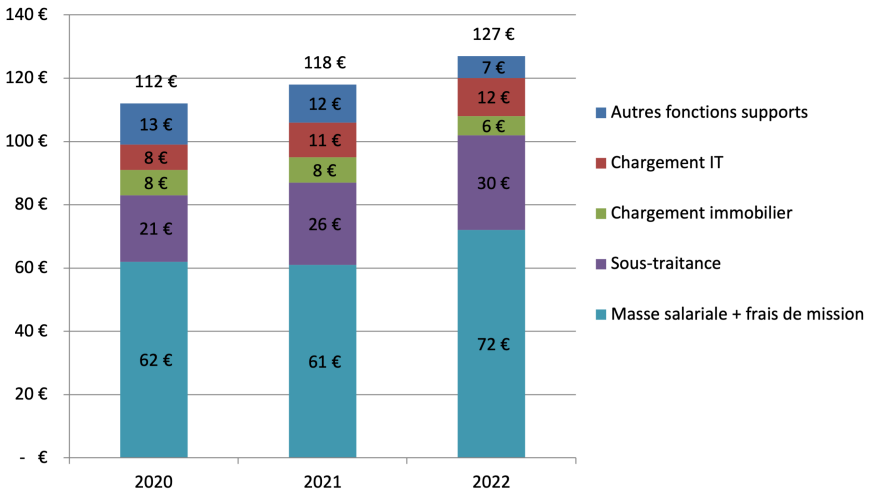


FIGURE 1.2 : Évolution et constitution du coût moyen de gestion

Il est possible d'y voir que le coût de gestion d'un sinistre connaît une tendance à la hausse, avec plusieurs facteurs clés contribuant à cette augmentation. Tout d'abord, la masse salariale représente la part la plus importante de ces coûts mais aussi de leur augmentation. De plus, on observe une progression du poids de la sous-traitance dans les coûts de gestion. La MAAF externalise de plus en plus certaines tâches liées aux sinistres notamment celles de prises d'appels. Cette tendance à la sous-traitance s'explique par la variabilité de la charge de gestion à laquelle doivent faire face les centres de gestion notamment lors des épisodes climatiques. La sous-traitance permet donc d'absorber ces

variations en plus de combler la réduction de capacité en période de vacances (notamment estivales).

Bien que la problématique de ventilation des coûts de gestion à l'échelle du dossier se pose pour le portefeuille de dossiers analysés pour les coûts des dégâts, il est possible de réaliser l'identification du dossier concerné dans certains cas. On peut donc représenter la répartition de ces coûts détectés. Pour les 22% des dossiers pour lesquels des coûts de gestion sont identifiés la distribution des coûts est donnée dans la figure 1.3.

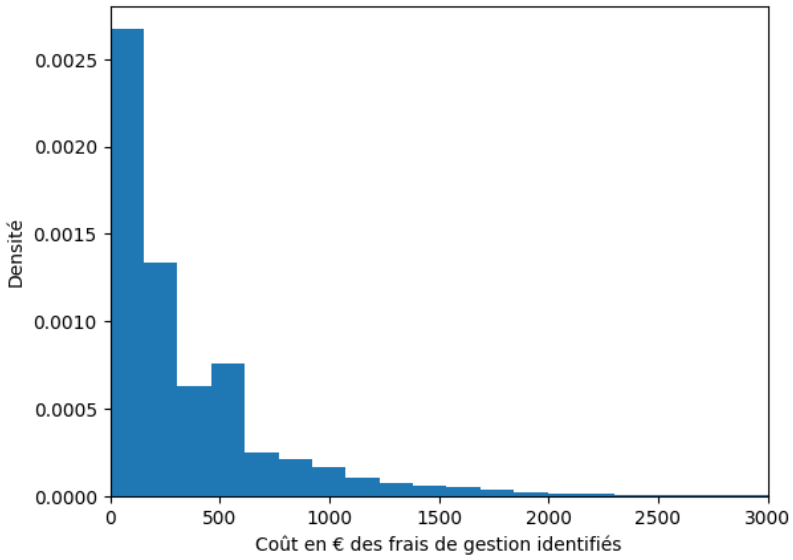


FIGURE 1.3 : Distribution des coûts de gestion identifiés

La moyenne de ces coûts de gestion identifiés (y compris sans suite et sans suite indemnitaire) est de $\mu_{gestion} = 90$ € et le coefficient de variation de $CV_{gestion} = 3,69 > 2,37 = CV_{dégâts}$. Il est à noter que le coût de gestion moyen réel sur ce portefeuille (obtenu via des données agrégées) est de 123 €.

En constatant que le coefficient de variation des coûts de gestion est supérieur à celui des coûts des sinistres, il apparaît opportun d'effectuer des études approfondies pour comprendre et gérer ces variations. L'analyse du processus de gestion et d'indemnisation devrait permettre de détecter les tendances, les inefficacités et les opportunités d'optimisation. En travaillant sur les facteurs qui contribuent à la volatilité des coûts de gestion, les actuaires peuvent aider les compagnies d'assurance à prendre des mesures appropriées pour réduire les écarts, améliorer l'efficacité opérationnelle et optimiser leur rentabilité. Adopter une approche actuarielle sur ces coûts de gestion semble indispensable pour une maîtrise efficace des coûts et une prise de décision éclairée dans le secteur de l'assurance.

La prépondérance de la masse salariale dans les coûts de gestion (voir figure 1.2) et le fait que la prise d'appel (de déclaration et de gestion) constitue plus de la moitié de l'activité d'un gestionnaire ont conduit à s'intéresser de manière plus approfondie aux appels.

1.2 Les flux téléphoniques, variable à forts enjeux

S'il a été décidé de focaliser l'étude sur les échanges téléphoniques c'est qu'ils sont au croisement de plusieurs objectifs. En effet, mieux comprendre la survenance des appels permettrait à la fois

d'améliorer la visibilité et la maîtrise des coûts de gestion mais également de renforcer la joignabilité et la satisfaction client.

1.2.1 Coût d'un appel de gestion

Afin de mieux saisir la part que représentent les appels dans le coût de gestion des sinistres il sera utile d'avoir à disposition un ordre de grandeur de ce que coûte un appel entrant à la MAAF. On sait à dires d'experts et grâce aux données téléphoniques qu'un appel occupe en moyenne 15 minutes du temps d'un conseiller. Si l'on ramène cette durée au coût global d'un équivalent temps plein d'un conseiller on peut évaluer le coût d'un appel à une quinzaine d'euros. Cet ordre de grandeur de 15 € est à mettre en rapport avec le coût moyen de gestion d'un sinistre sur la période 2021-2022 qui est de 123 € (on rappelle que l'on parle de sinistres de fréquence).

À ce coût de gestion d'un appel il faut rajouter le fait que ce soit en partie l'incertitude sur le volume d'appels entrants qui a conduit à une politique de sous-traitance onéreuse (voir figure 1.2).

1.2.2 La satisfaction client

Les appels jouent également un rôle majeur dans la satisfaction client et ce à plusieurs titres. Une bonne estimation de la charge d'appels entrants doit permettre un bon dimensionnement des capacités de prise d'appel. En effet, la joignabilité joue un rôle essentiel dans la satisfaction client et ce constat est d'autant plus vrai dans le secteur de l'assurance car l'appel intervient dans un moment potentiellement anxiogène pour l'assuré. Comme on peut le voir dans la figure 1.4 la joignabilité s'est fortement améliorée à la MAAF pour les dossiers IRD PRI.

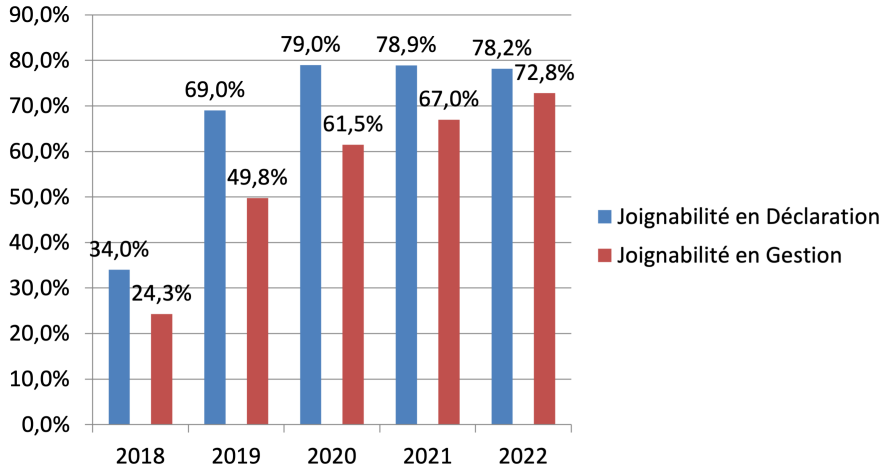


FIGURE 1.4 : Évolution de la joignabilité entre 2018 et 2022

Bien que la joignabilité s'améliore cette amélioration semble ralentir notamment en déclaration ces dernières années. Il est à noter que l'objectif de l'entreprise est d'obtenir une joignabilité de 80% que ce soit en déclaration ou en gestion.

Si la non-prise d'un appel dégrade l'expérience-client c'est également le cas la répétition des appels de gestion. En effet, lorsqu'un assuré contacte la MAAF après sa déclaration de sinistre il est à l'heure actuelle impossible de relier cet appel à son gestionnaire attitré ce qui peut être irritant. En pratique

le gestionnaire prenant l'appel devra relire le dossier et éventuellement poser des questions auxquelles l'assuré avait déjà répondu ce qui vient nuire à la satisfaction-client.

La raison pour laquelle la satisfaction client représente un enjeu majeur est qu'une dégradation de cette dernière, et ce tout particulièrement au moment de la gestion d'un sinistre, peut amener à des résiliations ou des non-renouvellements de contrat ce qui constitue un risque pour l'assureur. Il faut bien entendu rajouter le risque d'une dégradation de l'image d'entreprise qui peut pâtir d'une expérience client négative.

1.2.3 Problématique, besoin de compréhension du phénomène d'appel

Les flux téléphoniques se trouvant au croisement de plusieurs risques (coût de gestion et risques liés à l'expérience-client) il devient clair qu'ils doivent faire l'objet d'études plus approfondies. Partant de ce constat la suite de ce mémoire sera consacrée à comprendre et modéliser le déroulé du processus d'indemnisation et de gestion en mettant l'accent sur les flux téléphoniques subis par l'assureur.

1.3 Cadre, objectifs et données de l'étude

1.3.1 Cadre de l'étude

Notre étude portera sur des dossiers de sinistres IRD PRI ouverts en 2021 et 2022 à la MAAF. Seront exclus de l'étude les dossiers ayant fait appel à de la sous-traitance téléphonique car il est compliqué d'aller observer l'entièreté du processus de gestion pour ces derniers. Seront également exclus les dossiers pour lesquels des dommages corporels sont présents ou ceux dont les dégâts sont supérieurs à 23 000 € afin d'avoir un portefeuille cohérent. En effet, dans ces deux cas le dossier suivra un processus de gestion différent réalisé par des équipes dédiées.

1.3.2 Objectifs et méthodologie

Dans un premier temps l'objectif de l'étude sera de modéliser la survenance d'un appel de gestion (à l'initiative de l'assuré) dans un futur proche à l'échelle du dossier. Cette modélisation devra prendre en compte les caractéristiques du dossier ainsi que l'historique des actions connues ayant eu lieu par le passé. Par la suite, plutôt que de modéliser directement la survenance d'un appel de gestion une approche plus globale sera étudiée pour modéliser l'évolution d'un dossier. Le choix de se focaliser sur les appels de gestion est dû au fait que les appels de déclaration surviennent avant que l'assureur n'ait connaissance de l'existence du dossier or l'objectif est ici de faire une prédiction à l'échelle du dossier.

1.3.3 Données à disposition

Les dossiers de sinistres concernés et leurs caractéristiques

Les dossiers décrits ci-dessus représentent un portefeuille d'environ 400 000 dossiers ouverts entre 2021 et 2022. Parmi les principales caractéristiques de ces dossiers se trouvent la typologie de sinistre, le mode de gestion ou encore le canal de déclaration.

La typologie de sinistre, même sur le périmètre restreint de notre étude, reste variée et demeure une caractéristique essentielle des dossiers. En effet, le portefeuille portant quasi-exclusivement sur des contrats multirisques habitation (MRH) il est possible d'observer une large gamme de typologies de sinistre comme le montre la figure 1.5. Cette dernière montre que si les dégâts des eaux représentent un tiers des déclarations le portefeuille comporte également une part conséquente d'événements climatiques ainsi qu'une variété de typologies caractéristique des contrats MRH.

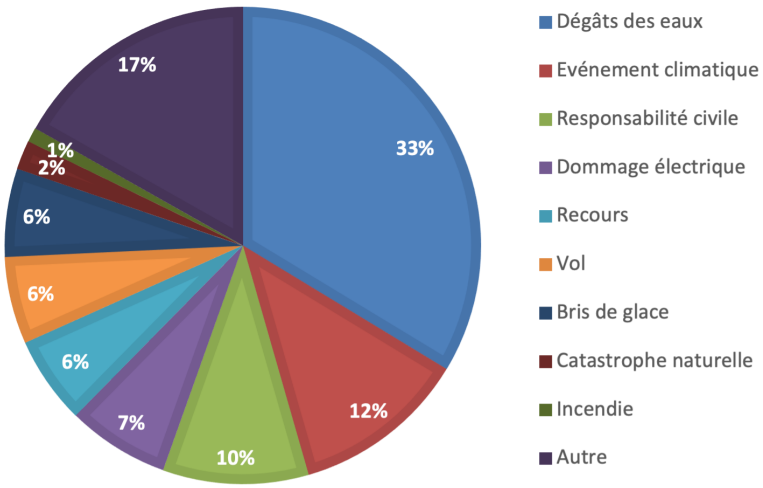


FIGURE 1.5 : Proportions des différentes typologies de sinistre du portefeuille de l'étude

Le mode de gestion est une variable stratégique pour l'assureur. Les modes les moins coûteux (à savoir le gré-à-gré, la réparation en nature et l'indemnisation sur devis facture) ne font pas intervenir d'expertise et donc sont plus sujets à la non-détection de fraude. Le fait de valoriser le gré-à-gré peut aussi entraîner une dérive en ouvrant des garanties à des dossiers qui devraient être classés sans suite. La figure 1.6 montre que pour les dossiers qui ne sont pas classés sans suite le règlement sur devis-facture reste l'option la plus répandue. Il est à noter que mis à part les dossiers sans suite, ceux gérés en gré-à-gré et ceux gérés par devis-facture tous les autres modes de gestion nécessitent l'intervention d'un autre professionnel en plus du gestionnaire.

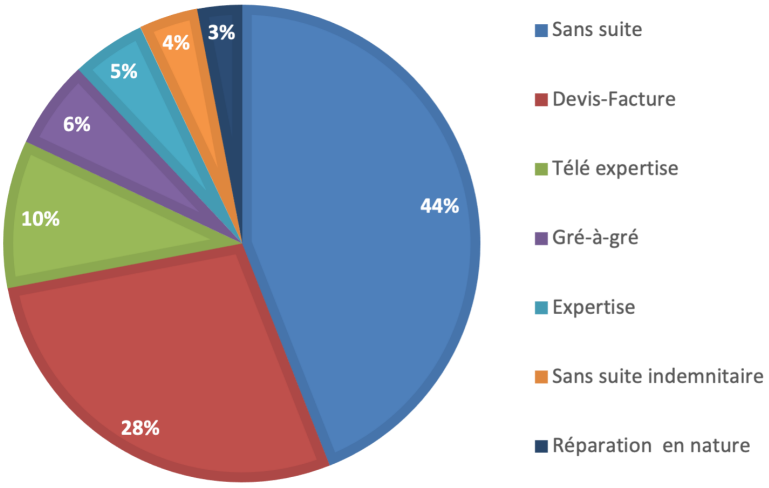


FIGURE 1.6 : Proportions des différents modes de gestion du portefeuille de l'étude

Le mode de contact à la déclaration est également une donnée intéressante car d'une part c'est la première information qui parvient à l'assureur à l'ouverture du dossier mais aussi car il est le fruit du

choix de l'assuré et peut donc traduire ses préférences notamment pour l'utilisation d'outils digitaux. On peut voir sur la figure 1.7 que le téléphone reste le mode de contact le plus choisi pour la déclaration et que les outils digitaux (regroupés dans la catégorie Internet) restent encore peu populaires.

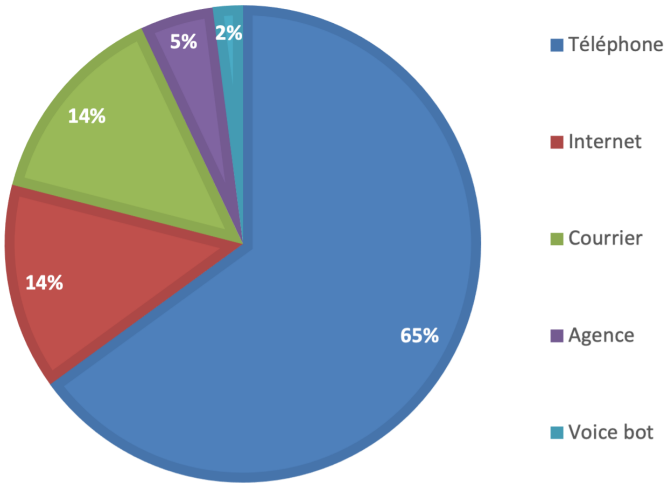


FIGURE 1.7 : Proportions des différents modes de contact à la déclaration du portefeuille de l'étude

Les flux et événements

En plus des caractéristiques évoquées ci-dessus les dossiers possèdent un historique d'événements qui vient grandir tout au long de la gestion. Ces événements sont tous matérialisés par des flux. Ainsi l'historique d'un dossier est composé d'événements appartenant à la liste suivante :

- Ouverture du dossier dans le système informatique
- Appel de déclaration
- Appel sortant
- Appel de gestion
- Pli (manuscrit ou digital) reçu
- Pli (manuscrit ou digital) émis
- Ouverture d'un ordre de mission d'expertise ou de réparation
- Réalisation de la prestation (versement ou réparation)
- Clôture du dossier

Pour chacun de ces événements au moins une date est disponible avec parfois des informations complémentaires. Comme rappelé ci-dessus, dans un premier temps l'objectif de l'étude sera de modéliser la survenance proche (dans un sens qui sera spécifié) de l'événement "Appel de gestion" en fonction des caractéristiques et de l'historique du dossier. En pratique, en se plaçant à une date donnée un certain nombre de dossiers sont ouverts mais avec des profondeurs d'historiques extrêmement variables. Cette variabilité s'explique d'une part car les dossiers ne sont pas tous ouverts depuis la même durée mais aussi car certains dossiers ont une gestion plus lourde que d'autres.

1.3.4 Focus sur les appels de gestions

Variable d'intérêt de cette étude les appels de gestion vont ici être analysés de manière descriptive afin de dégager des résultats d'ordre général afin de mieux cerner les phénomènes qu'une bonne modélisation devrait être capable de retranscrire.

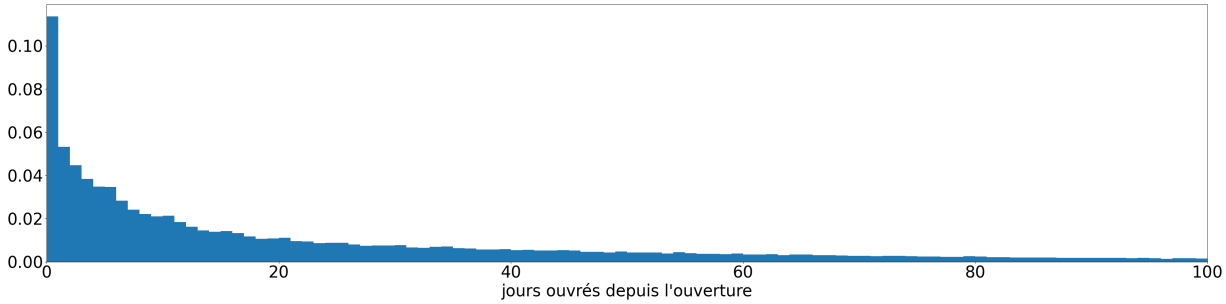
Problématique d'identification des appels

Avant de poursuivre l'analyse il est primordial de souligner que pour de nombreuses raisons il est possible qu'un appel de gestion ait eu lieu sans qu'il soit possible de l'identifier et le faire apparaître dans l'historique du dossier. En effet, pour qu'un appel puisse être relié à un dossier il faut que le numéro appelant corresponde à celui de la fiche client du client sinistré. Dans la pratique certains assurés peuvent appeler en numéro masqué, avec leur téléphone professionnel ou plus généralement avec un numéro non connu du système d'information. Dans certains cas une personne peut appeler pour gérer le dossier d'une autre (un enfant pour un parent âgé typiquement) et dans ce cas en plus de ne pas apparaître dans l'historique du dossier cet appel peut apparaître dans un autre dossier si la personne qui appelle est également assurée chez la MAAF avec un dossier de sinistre ouvert.

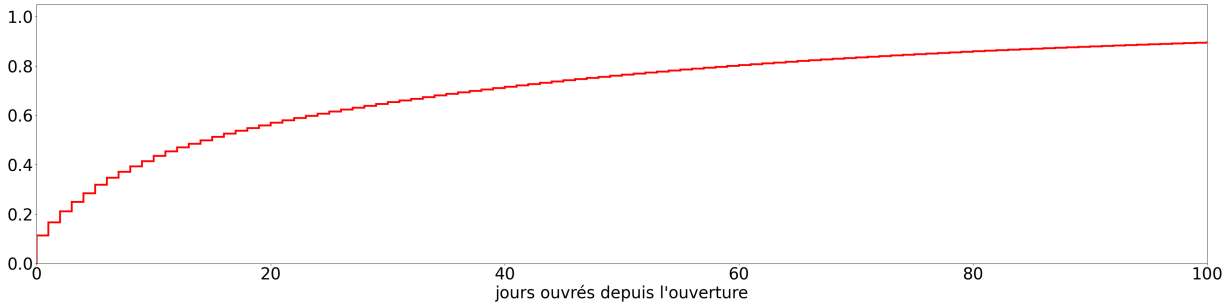
Dans ce qui va suivre il faut donc considérer les chiffres qui vont être exposés comme des bornes inférieures. Sur le portefeuille de notre étude il a été estimé qu'entre 55% et 65% des appels de gestion étaient reliés à un dossier.

Répartition des appels de gestion sur la vie d'un dossier

Dans les figures 1.8a et 1.8b il est possible de constater la répartition des appels de gestion sur la vie d'un dossier. Plus on est proche de l'ouverture du dossier plus l'activité téléphonique est intense. Cela est d'autant plus vrai pour le jour de l'ouverture qui concentre plus de 11% du total des appels de gestions.



(a) Histogramme de la distribution des appels de gestion depuis l'ouverture



(b) Fonction de répartition de la distribution des appels de gestion depuis l'ouverture

FIGURE 1.8 : Distribution des appels de gestion sur la vie du dossier

Cette très forte activité le jour même de la déclaration peut laisser perplexe surtout en sachant que la plupart des assurés ont fait leur déclaration par téléphone (voir 1.7). On est donc face à de nombreux cas où l'assuré rappellera le jour même après avoir déjà eu un gestionnaire au téléphone. Ce constat peut laisser penser que lors de la déclaration il y aurait une réalisation partielle des tâches qui devraient y être accomplies.

Au-delà de la répartition de l'ensemble des appels de gestion, il est intéressant de savoir combien de fois un assuré va appeler en gestion et à quelles dates vont survenir les différents appels. Les graphiques qui vont suivre doivent s'interpréter de la manière suivante, la n-ième barre représente en ordonnée la part des dossiers ayant au moins eu n appels et en abscisse la date médiane du n-ième appel. La figure 1.9 montre que 34% des dossiers ont au moins un appel (se rappeler qu'on ne compte ici que les appels identifiés) et que ce premier appel intervient en médiane au bout de 5 jours ouvrés. Ensuite 13% des dossiers ont au moins un deuxième appel qui intervient en médiane à 14 jours ouvrés. La dynamique semble être la suivante, les assurés qui appellent le font au bout d'une semaine puis sont à chaque appel deux fois moins nombreux et espacent les appels de 2 semaines.

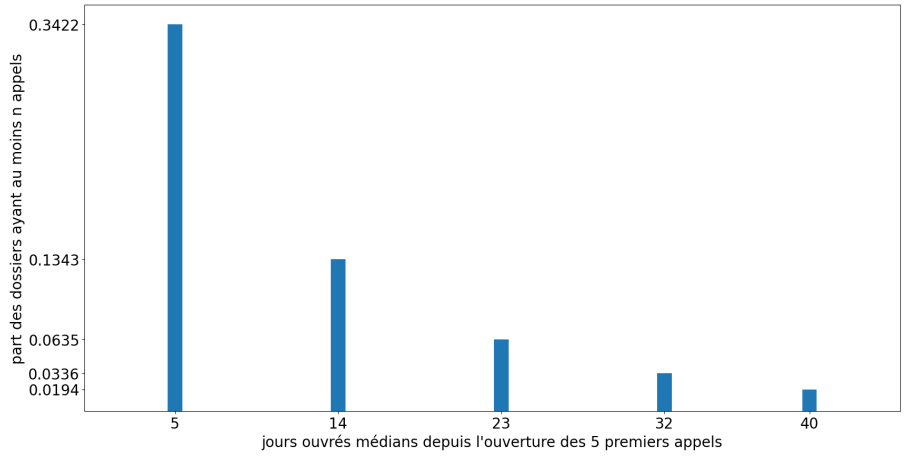
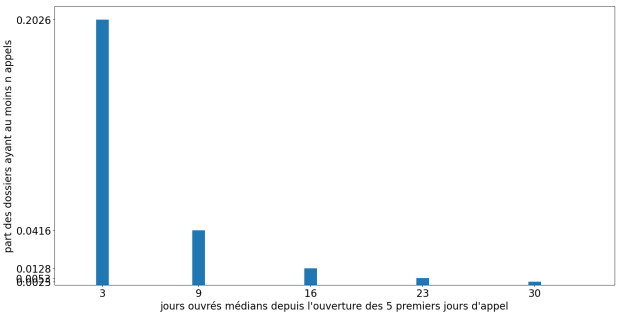
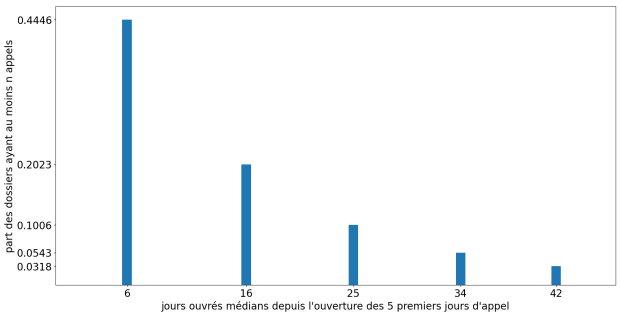


FIGURE 1.9 : Proportion des dossiers ayant au moins n appels et date médiane du n-ième appel

En séparant les dossiers sans-suite (figure 1.10a) des dossiers hors sans-suite (figure 1.10b) il apparaît que les dossiers sans-suite sont moins générateurs d'appels mais que ceux-ci interviennent plus tôt que dans les autres dossiers. Cette différence laisse entrevoir une différence de comportement de l'assuré en fonction du classement sans-suite ou non de son dossier de sinistre.



(a) Profil médian d'appel des sans-suite



(b) Profil médian d'appel des hors sans-suite

FIGURE 1.10 : Proportion des dossiers ayant au moins n appels et date médiane du n-ième appel

Segmentation selon les caractéristiques principales du dossier

En allant segmenter la répartition des appels selon les caractéristiques du dossier, il ressort que ces dernières permettent d'observer des volumes comme des répartitions de flux téléphoniques variés. Ce phénomène peut s'observer sur les figures 1.11, 1.12 et 1.13. Ces figures doivent s'interpréter de la manière suivante, à côté de la caractéristique est rappelée la proportion qu'elle occupe dans le portefeuille puis les différentes sections de couleurs représentent le nombre d'appels de gestion ayant eu lieu sur la période affectée. Les périodes sont données en jours ouvrés, ainsi la section bleue représente le jour de déclaration, les sections bleue orange et verte la première semaine de gestion et ainsi de suite jusqu'à la section rose qui correspond à la période comprenant le troisième mois et tous les suivants.

Ainsi la figure 1.11 révèle que les incendies génèrent le plus d'appels de gestion et que si l'on s'intéresse uniquement au jour de déclaration ce sont les événements climatiques qui sont les moins générateurs de ces flux. Les sinistres de vol présentent une forte activité en première semaine, comparable aux sinistres de dégâts des eaux, mais voient leur charge se réduire dans le temps.

Sur la figure 1.12 il s'observe que les modes de gestion faisant intervenir des professionnels sont les plus générateurs d'appels de gestion et que malgré sa réputation de mode de gestion plus simple le gré-à-gré engendre plus d'appels que le règlement sur devis-facture. À noter également que c'est sur le temps long que les modes de gestion les plus lourds acquièrent leur charge de gestion excédentaire.

Enfin, sur la figure 1.13 un constat surprenant se dégage, les assurés ayant fait leur déclaration par téléphone font partie de ceux qui appellent le plus en gestion le jour même alors qu'ils ont pu avoir au préalable un interlocuteur pour répondre à leurs interrogations. Autre fait qui interpelle, les assurés ayant fait le choix du canal internet pour la déclaration semble ensuite générer plus d'échanges téléphoniques.

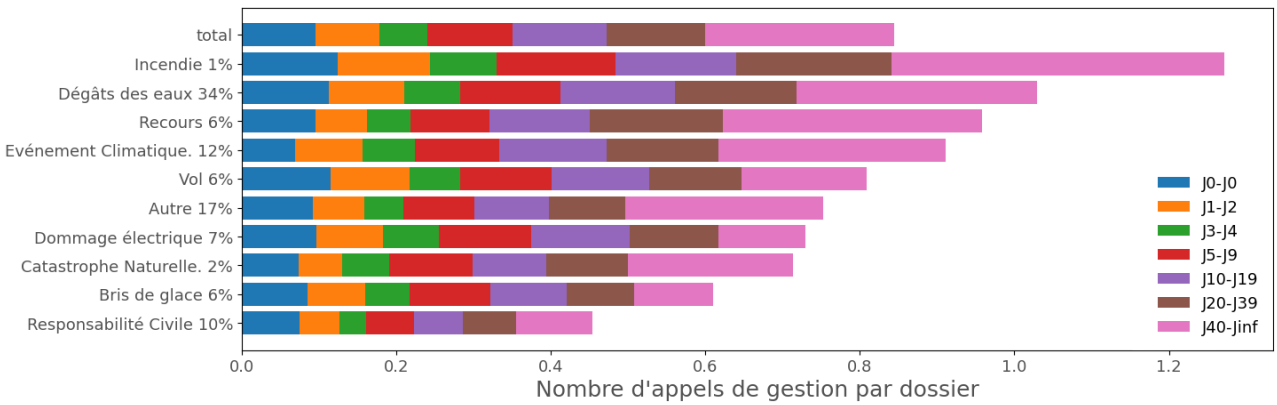


FIGURE 1.11 : Répartition des appels de gestion selon la typologie de sinistre

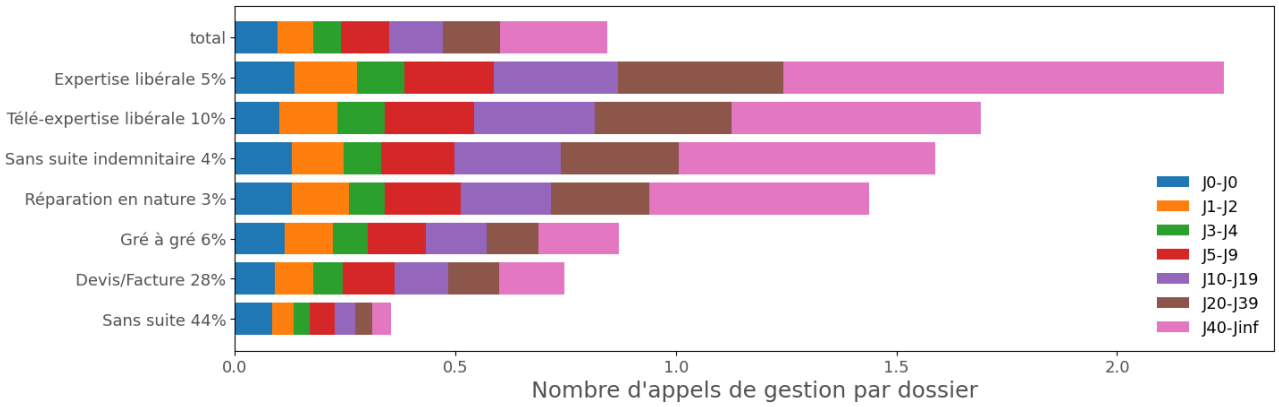


FIGURE 1.12 : Répartition des appels de gestion selon le mode de gestion

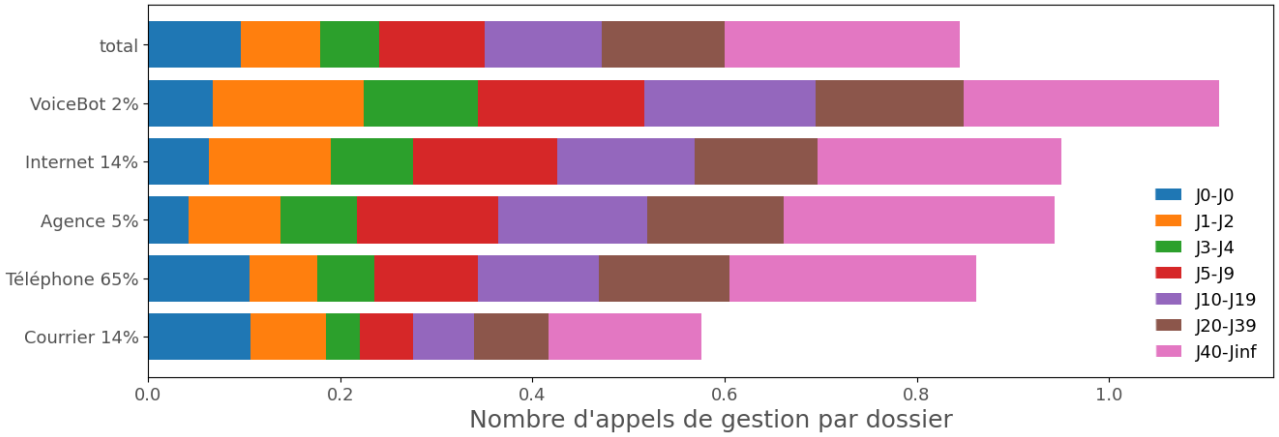


FIGURE 1.13 : Répartition des appels de gestion selon le mode de contact à la déclaration

Ces premières statistiques descriptives laissent transparaître des différences notables de comportements. En ajoutant la complexité de l'historique d'événements aux caractéristiques du dossier il devient clair que pour être mieux compris le phénomène d'appel doit faire l'objet d'une modélisation. Au vu des éléments développés dans ce premier chapitre le lecteur peut considérer le phénomène d'appel comme une forme de sous-sinistralité. En effet, l'appel de gestion intervient de manière aléatoire au sein du dossier de sinistre et vient en affecter le coût global. Mieux comprendre le phénomène doit donc permettre de mieux analyser une partie du risque pesant sur l'assureur.

Chapitre 2

Premières modélisations du phénomène d'appel via des outils de data science

Dans ce second chapitre il va être développée une stratégie pour modéliser l'occurrence future d'appels de gestion pour un dossier en cours de gestion. Les interprétations qui peuvent être tirées de cette modélisation seront également étudiées.

2.1 Objet de la modélisation, métriques d'évaluation

La modélisation retenue devra être à l'échelle du dossier, prendre en compte ses caractéristiques connues et son passé. La pertinence de la modélisation doit permettre d'identifier les dossiers par leur risque d'appel, il s'agit donc d'un problème d'affectation d'un score à chaque dossier.

2.1.1 Cadre mathématique

Le problème va être modéliser comme suit, on se place à une date t (en jours) et on s'intéresse à n'importe quel dossier d qui soit ouvert à la date t et on cherche à estimer la probabilité que le dossier d occasionne au moins un appel de gestion dans les deux semaines suivantes soit avant $t + 14$. Plus formellement, on cherche à estimer la probabilité suivante.

Définition 2.1 (Probabilité d'appel)**.** Soit t une date et d un dossier ouvert à la date t, soit AG_s^d le nombre total d'appels de gestion du dossier d intervenus avant s. On définit la probabilité d'appel du dossier d en t, notée a_t^d, par

$$\begin{aligned} a_t^d :&= \mathbb{P}(AG_{t+14}^d > AG_t^d \mid \mathcal{F}_t^d) \\ &= \mathbb{E}(\mathbb{1}_{AG_{t+14}^d > AG_t^d} \mid \mathcal{F}_t^d), \end{aligned}$$

où $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ est une filtration qui contient l'évolution de la connaissances sur les informations (caractéristiques et événements) du dossier d.

Remarque 2.1. Dans la suite la dépendance en t ou d des notations sera parfois omise pour simplifier les notations.

Dans la suite on travaillera donc sur le problème de classification suivant.

Définition 2.2. On note Y la variable dont on veut effectuer la classification binaire définie par

$$Y := \mathbb{1}_{AG_{t+14}^d > AG_t^d}.$$

Il sera considéré que $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ regroupe les informations suivantes :

- la date d'ouverture du sinistre
- la typologie de sinistre
- le mode de gestion
- le mode de contact à la déclaration
- les dates et natures des événements connus
- savoir si la prestation a été réalisée

On considère donc que dès l'instant de l'ouverture, la date d'ouverture, la typologie de sinistre, le mode de gestion et le mode de contact sont connus ce qui semble cohérent sauf éventuellement pour le mode de gestion. En effet, ce dernier peut être déterminé ultérieurement ou être amené à évoluer. Cependant, afin d'analyser également l'impact du mode de gestion sur la survenance d'appel, cette information sera considérée comme toujours connue quitte à devoir traiter une problématique de valeurs manquantes en pratique.

S'il a été choisi de s'intéresser à la probabilité de survenance d'au moins un appel dans les deux semaines plutôt qu'à l'espérance du nombre d'appels sur la même période c'est que l'étude doit permettre d'identifier le besoin de l'assuré pour pouvoir éventuellement être proactif et éviter l'appel. C'est pour la même raison que les appels non-traités seront également pris en compte toujours dans le but de cibler le besoin de l'assuré.

2.1.2 Choix des métriques (AUC, nombres d'appels détectés)

Afin d'évaluer la qualité d'un estimateur de a_t^d mais aussi de comparer différents estimateurs entre eux il est nécessaire de disposer d'au moins une métrique de performance. Il s'agit donc d'avoir une métrique de performance pour un classifieur binaire qui renvoie une valeur réelle (et non directement une valeur binaire).

Courbe ROC et AUC

Si l'on dispose d'un estimateur de a_t^d, on dispose en réalité d'une infinité d'estimateurs si l'on considère le problème de classification binaire sur la variable Y (présence ou non d'un appel avant deux semaines). Un estimateur $\widehat{a}$ de a étant à valeurs réelles, pour un seuil p on obtient un classifieur binaire qui conclut à la présence ou non d'un appel à moins de deux semaines si et seulement si $\widehat{a} > p$.

En fixant un seuil p on dispose d'un estimateur $\widehat{Y_p}$ de Y qu'on peut appliquer sur tous les dossiers ouverts en t et on peut donc construire une matrice de confusion comme dans le tableau 2.1.

		Y	
		1	0
$\widehat{Y_p}$	1	Vrai Positif (VP)	Faux Positif (FP)
	0	Faux Négatif (FN)	Vrai Négatif (VN)

TABLE 2.1 : Matrice de confusion

Si on note D_t l'ensemble des dossiers ouverts en t on a donc :

$$VP = \sum_{d \in D_t} \mathbb{1}_{\widehat{Y_p}=1, Y=1}, \tag{2.1}$$

$$FP = \sum_{d \in D_t} \mathbb{1}_{\widehat{Y_p}=1, Y=0}, \tag{2.2}$$

$$FN = \sum_{d \in D_t} \mathbb{1}_{\widehat{Y_p}=0, Y=1}, \tag{2.3}$$

$$VN = \sum_{d \in D_t} \mathbb{1}_{\widehat{Y_p}=0, Y=0}. \tag{2.4}$$

À partir de la matrice de confusion on peut déterminer de nombreuses métriques dont le taux de vrais positifs (TVP) et le taux de faux positifs (TFP).

Définition 2.3. Le taux de vrais positifs (TVP) et le taux de faux positifs (TFP) sont définis comme

$$TVP := \frac{VP}{VP + FN}, \tag{2.5}$$

$$TFP := \frac{FP}{FP + VN}. \tag{2.6}$$

Remarque 2.2. TVP et TFP sont croissant en le seuil p car VP et FP le sont tandis que $VP + FN$ et $FP + VN$ sont constants.

Ces deux quantités s'interprètent comme le bénéfice (TVP) et le coût (TFP) d'un estimateur dans le sens où elles indiquent quelle proportion des négatifs il a fallu retenir comme positifs pour obtenir une proportion de positifs bien classés. C'est ce principe combiné au fait que l'on dispose d'un estimateur par seuil p qui permet de définir la courbe ROC.

Définition 2.4. La courbe ROC (Receiver Operating Characteristic) d'un estimateur de probabilité de la classe positive est définie comme l'ensemble des point de coordonnées (TFP, TVP) pour chaque seuil possible.

Dans l'espace où vit la courbe ROC chaque point correspond à la performance d'un classifieur binaire. Le point $(0, 1)$ correspond à l'estimateur parfait (capable de détecter tous les positifs sans avoir besoin de prendre de négatif). Le point $(0, 0)$ (respectivement $(1, 1)$) correspond à l'estimateur qui classifie tout le temps en négatif (respectivement positif).

Comme l'illustre la figure 2.1 un estimateur aléatoire de probabilité donne une courbe ROC égale à l'identité, tandis qu'un estimateur qui permet de séparer parfaitement les classes donne une courbe ROC qui relie $(0, 0)$ à $(0, 1)$ et $(0, 1)$ à $(1, 1)$. Plus une courbe ROC est proche de celle de l'estimateur parfait plus le modèle qu'elle représente est performant ce qui conduit à l'utilisation de l'AUC comme mesure de performance.

Définition 2.5. Soit un modèle qui estime la probabilité de la classe positive dans un problème de classification binaire, l'AUC (Area Under the Curve) de ce modèle est définie comme l'air sous sa courbe ROC

Remarque 2.3. Un modèle parfait a donc une AUC de 1 tandis qu'un modèle aléatoire aura une AUC de 0,5.

Dans la suite l'AUC sera notre principal critère de sélection pour comparer et évaluer des estimateurs de la probabilité d'appel.

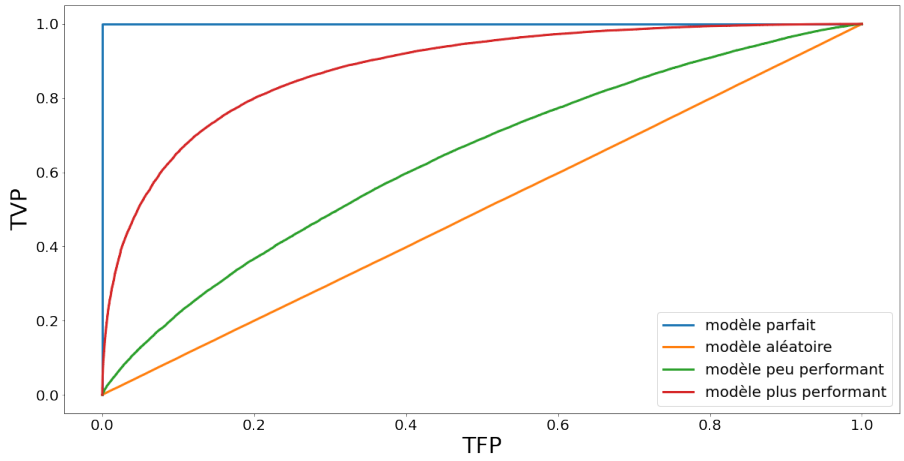


FIGURE 2.1 : Courbes ROC de différents modèles pour un problème de classification binaire

Nombre d'appels détectés à capacité donnée

Comme la prédiction sur la probabilité d'appel va également être utilisée dans une optique de proactivité avec priorisation des dossiers et que les capacités affectées à cette tâche seront limitées, il est également important que le modèle identifie des dossiers à "forte" probabilité d'appel. En pratique il sera considéré qu'un fois les probabilités estimés les gestionnaires pourront a minima prioriser 500 dossiers et un enjeu est donc de pouvoir sélectionner 500 dossiers de manière à maximiser ceux qui auraient effectivement appelé si leur dossier avait continué à être géré de manière classique. On cherche donc à évaluer la précision d'un modèle uniquement sur ses hautes prédictions. Cela conduit à introduire la métrique simple suivante.

Définition 2.6. Soit une date t, D_t l'ensemble des dossiers ouverts en t et $\left(\widehat{a}_t^d\right)_{d\in D_t}$ les valeurs d'un estimateur de la probabilité d'appel. On définit le nombre d'appels détectés à capacité n donné comme suit

$$DETEC_t(\widehat{a}, n) := \sum_{i=1}^{n} \mathbb{1}_{AG_{t+14}^{d_i} > AG_t^{d_i}},$$

où d_i est le dossier ayant la i-ème plus grande valeur dans $\left(\widehat{a}_t^d\right)_{d\in D_t}$.

En réalité, il serait plus exact de parler de nombre de dossiers appelants plutôt que d'appels. Par la suite en plus de l'AUC, $DETEC_t(\cdot, 500)$ sera donc également utilisé comme métrique d'évaluation pour les modèles.

2.2 Mise en forme d'une base mixte événements/caractéristiques

Afin de pouvoir utiliser des outils de data science il faut déterminer une façon de représenter la filtration $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ sous la forme d'une variable dans $\mathbb{R}^n$, $n \in \mathbb{N}^*$ et un estimateur de la probabilité d'appel sera donc une fonction de cette variable.

2.2.1 Problématique de structuration pour les outils classiques

En effet, la plupart des modèles, notamment les plus simples et interprétables, reposent sur le problème théorique suivant. On cherche à approcher un variable aléatoire Y grâce à l'information

apportée par une variable aléatoire $X \in \mathbb{R}^n, n \in \mathbb{N}^*$ et donc dans le cas où Y est une variable aléatoire réelle intégrable à approcher la fonction f telle que $\mathbb{E}(Y \mid X) = f(X)$. Cela se rapproche donc du problème d'estimation de la probabilité d'appel (voir définition 2.1) à supposer que l'on remplace la filtration $\left(\mathcal{F}_t^d\right)_{t\geq 0}$ par $X \in \mathbb{R}^n$.

Étant donné qu'il a été supposé que la filtration regroupe les informations sur certaines variables une solution intuitive serait de simplement concaténer ces variables pour créer une variable dans $\mathbb{R}^n$. Cependant, il n'est pas possible de procéder ainsi pour deux raisons principales. Premièrement, certaines variables sont catégorielles (typologie de sinistre, mode de gestion et canal de déclaration) mais il existe des solutions pour transformer des variables catégorielles en variables numériques. Deuxièmement, les événements sont présents en nombre variable pour différents dossiers ce qui est un obstacle à la construction d'un encodage identique pour représenter les événements dans la même dimension pour chaque dossier.

Au-delà de cette problématique de nombre d'événements variable, la façon de représenter une date au sein d'un dossier est un choix déterminant. Il est possible de fixer une date de référence et d'observer toutes les dates par leurs distances à cette référence mais le modèle devant se placer dans une vision opérationnelle il doit être capable de s'appliquer à différentes dates et une référence unique viendrait rendre incohérente cette utilisation en "temps continu".

2.2.2 Principe de scan

C'est cette perspective d'utilisation sur des données arrivant en temps réel qui contraint la construction d'un modèle (sur un jeu de données passé) à se faire selon le principe d'un scan des données. Il faut pour une date donnée en plus de savoir quels dossiers étaient ouverts à ce moment-là, sélectionner les données qui étaient effectivement disponibles. Cette opération est indispensable pour construire un ensemble d'entraînement cohérent et ne pas construire un modèle qui serait performant sur des données historiques mais qui utiliserait des informations non-connues à date en pratique.

Pour la plupart des événements il suffit d'observer s'ils ont eu lieu avant la date de scan car ils sont enregistrés dans les systèmes d'information de manière automatique (relevé téléphonique, envoi de plis, etc.) en revanche la réception de pli doit être considérée avec soin car deux dates sont disponibles pour chaque réception de pli, une date de réception et une date de classement. Bien que l'action du gestionnaire ne soit réalisée qu'au moment du classement, on souhaite se placer du point de vue de l'assuré car on cherche à prédire son comportement. Ainsi dans tout ce qui suit la date d'un pli reçu devra être considérée comme celle de sa réception.

Selon ce principe de scan des informations connues à date il est possible de disposer d'un ensemble de données arbitrairement grand en se plaçant à plusieurs dates de scan. Un dossier peut être ouvert à différentes dates de scan et donc regarder plusieurs fois dans les données mais possiblement avec des informations connues différentes.

Dans la mesure où la variable à expliquer est construite sur les deux semaines suivant la date de scan, il faut se restreindre en pratique à des dates de scan vieilles d'au moins deux semaines afin de pouvoir observer la variable à expliquer et donc d'étiqueter correctement les données quant à la présence ou non d'un appel de gestion dans les deux semaines suivantes. Le choix des dates de scan doit aussi se faire de manière cohérente pour avoir un ensemble d'entraînement semblable aux conditions dans lesquelles le modèle sera évalué.

2.2.3 Solution retenue

Afin d'illustrer la solution retenue, il est important de visualiser comment se présente la base des événements ce qui est possible avec le tableau 2.2. Il peut exister des exemples de dossiers bien plus longs en gestion que ce soit en termes de durée ou de nombre d'événements que les trois exemples simples présentés.

Numéro de dossier	date	événement
1	11/02/2022	ouverture
1	11/02/2022	appel de déclaration
1	15/02/2022	pli reçu
1	15/02/2022	pli envoyé
1	17/02/2022	appel sortant
1	22/02/2022	appel de gestion
1	23/02/2022	appel de gestion
1	01/03/2022	pli envoyé
1	01/03/2022	clôture
2	14/02/2022	ouverture
2	14/02/2022	appel de déclaration
2	24/02/2022	pli envoyé
2	25/02/2022	appel sortant
2	25/02/2022	pli envoyé
2	25/02/2022	clôture
3	17/02/2022	ouverture
3	17/02/2022	pli reçu
3	18/02/2022	appel de gestion
3	19/02/2022	appel sortant
3	20/02/2022	mission
3	04/04/2022	pli reçu
3	04/04/2022	clôture

TABLE 2.2 : Exemple d'événements pour 3 dossiers

On dispose également d'une base de caractéristiques (que l'on suppose connues dès l'ouverture) qui se présente comme dans le tableau 2.3. Le canal de déclaration correspond bien à celui observable au tableau 2.2.

Numéro de dossier	typologie (typo)	mode de gestion (MG)	canal de déclaration (CD)
1	incendie	devis-facture	téléphone
2	dégâts des eaux	gré-à-gré	téléphone
3	événement climatique	expertise	courrier

TABLE 2.3 : Exemple de caractéristiques pour 3 dossiers

En plus de ces deux bases de données on dispose d'un historique permettant de savoir à quelle date la prestation a été réalisée qui permet de déterminer si à une date un dossier est en Intérêt Client ou non. Un dossier est dit en Intérêt Client si la prestation n'a pas encore été réalisée. Cette approche par Intérêt Client plutôt que de considérer la réalisation de la prestation comme un événement a pour but de bien rendre identifiables les dossiers qui sont encore dans l'attente d'une prestation.

Avant de traiter le problème des variables catégorielles, la façon de représenter l'historique des événements va être fixée. En partant d'une représentation où un même dossier comporte plusieurs lignes d'événements (voir tableau 2.2) on se ramène à une représentation standardisée en utilisant une profondeur donnée.

Définition 2.7 (Scan de profondeur donnée)**.** Soient t une date, d un dossier ouvert en t et $n \in \mathbb{N}$. Soit $\left(e_i^d, t_i^d\right)_{i\in\{1,\dots k_d\}}$ l'ensemble des couples $(\textit{événement}, \textit{date})$ du dossier d classé du plus ancien au plus récent avec k_d le nombre total d'événements du dossier d. On a $e_i^d \in$ {ouverture, appel de déclaration, appel de gestion, appel sortant, pli reçu, pli envoyé, mission, clôture, indéfini} cette dernière valeur correspondant au cas $i \leq 0$. On définit le scan du dossier d de profondeur n à la date t noté $SCAN(t, d, n)$ par

$$SCAN(t, d, n) = \left(e_i^d, \delta_i^d\right)_{i\in\{k_d(t)-n+1,\dots,k_d(t)\}}, \tag{2.7}$$

$$\text{où } \delta_i^d = \begin{cases} t - t_{k_d}^d(t) & \text{si } i = k_d(t) \\ t_{i+1}^d - t_i^d & \text{si } 1 \leq i < k_d(t) \\ 0 & \text{sinon} \end{cases}, \tag{2.8}$$

$$\text{et } e_i^d = \text{ indéfini si } i \leq 0 \ , \tag{2.9}$$

avec $k_d(t) = \sum_{i=1}^{k_d} \mathbb{1}_{t_i^d<t}$ le nombre d'événements survenus avant t.

En d'autres termes on regarde les n derniers événement connus en remplaçant éventuellement par une valeur par défaut lorsque que l'historique est moins profond que n. L'avantage de calculer δ_i^d comme la distance à la date de l'événement suivant ou la date de réalisation du scan pour le plus récent est de pouvoir à la fois comparer des données scannées à différentes dates mais aussi de mieux prendre en compte la proximité entre les événements. En effet, si par exemple les δ_i^d étaient tous calculés avec la date t de scan comme référence on observerait plus difficilement le fait que deux dossiers aient eu une activité similaire mais à différents moments.

Le tableau 2.4 représente un scan de profondeur 3 à la date du 22/02/2022 pour les trois dossiers présents dans le tableau 2.2 afin que le lecteur puisse se représenter la définition 2.7. On peut bien y observer que de cette manière les événements de chaque dossier ont été résumés dans un format unique.

N° de dossier	$\delta_{k_d(t)}$	$e_{k_d(t)}$	$\delta_{k_d(t)-1}$	$e_{k_d(t)-1}$	$\delta_{k_d(t)-2}$	$e_{k_d(t)-2}$
1	5	appel sortant	2	pli envoyé	0	pli reçu
2	8	appel de déclaration	0	ouverture	0	indéfini
3	2	mission	1	appel sortant	1	appel de gestion

TABLE 2.4 : Exemple de scan de profondeur 3 pour 3 dossiers

Par la suite, s'il n'y a pas d'ambiguïté, on notera $\delta, e, \delta_{-1}, e_{-1}, \delta_{-2}, e_{-2}, \dots$ pour parler des variables issues du scan de profondeur donnée pour simplifier les notations. En pratique on utilisera un scan de profondeur 6, cette valeur étant issue d'un arbitrage entre informations représentées et nombre de variables. Si l'on concatène les variables issues du scan à celles des caractéristiques et qu'on rajoute une variable binaire indiquant si le dossier est en Intérêt Client (notée IC) ainsi qu'une variable indiquant depuis combien de jours le dossier est ouvert au moment du scan (notée δ_{ouv}), on obtient un ensemble de variables qu'on considère bien approcher la filtration $\left(\mathcal{F}_t^d\right)_{t\geq 0}$. L'ensemble des variables qui serviront de variables prédictives est données au tableau 2.5 en indiquant si elle sont catégorielles (cat), numériques (num) ou binaires (dans {0,1}). Il reste à transformer les variables catégorielles en variables réelles. On réalise cette transformation au moyen d'un encodage one-hot.

Variables	δ	e	δ_{-1}	e_{-1}	...	δ_{-5}	e_{-5}	δ_{ouv}	typo	MG	CD	IC
Type	num	cat	num	cat	...	num	cat	num	cat	cat	cat	binaire

TABLE 2.5 : Variables prédictives

Définition 2.8 (Encodage one-hot)**.** Soit C une variable aléatoire catégorielle à valeurs dans un espace d'états fini non ordonné $V = \{v_1, v_2, ..., v_n\}$, $n \in N^*$. On définit l'encodage one-hot de C comme la variable aléatoire à valeur dans $\mathbb{R}^n$ où la i-ème coordonnée est $\mathbb{1}_{C=v_i}$.

Il est obligatoire de traiter chaque valeur comme une variable car dans notre cas il n'existe pas de lien hiérarchique quelconque qui permettrait d'instaurer une relation d'ordre et de créer une seule variable réelle par variable catégorielle. On ne s'intéresse pas non plus aux interactions entre les différentes variables catégorielles car si avec un encodage one-hot on crée autant de variables que la somme des dimensions des espaces d'états il faut prendre le produit si on tient compte des interactions. En ayant un scan de profondeur 6, 8 événements possibles, 10 modes de gestion (certains très spécifiques n'ont pas été présentés), 15 typologies de sinistres et 11 canaux de déclaration un encodage one-hot aboutit à une dimension $6 \times 8 + 10 + 15 + 11 = 84$ ce qui reste acceptable pour la plupart des modèles avec une bonne capacité de calcul. En tenant compte des interactions il faudrait une dimension $8^6 \times 10 \times 15 \times 11 = 432\ 537\ 600$ ce qui serait rédhibitoire à l'utilisation.

Il est possible de réduire d'une dimension l'encodage one-hot de chaque variable catégorielle, la somme de toutes les coordonnées étant constante égale à 1. Ce choix ne sera pas retenu pour deux raisons. D'une part, avoir une variable par état possible permet une meilleure interprétabilité du modèle. D'autre part, certains modèles, par leur fonctionnement, sont capables de capter des interactions, il est donc préférable que les interactions soient observables sans avoir à combiner toutes les variables.

2.2.4 Ensemble d'entraînement et de test

Avec l'application de l'encodage one-hot on dispose d'une procédure permettant d'obtenir un jeu de variables prédictives (91 au total) pour une date donnée. Si cette date se situe au moins deux semaines dans le passé on peut aussi observer la variable cible (présence ou non d'un appel de gestion). Il reste maintenant à déterminer sur quels jeux de données les modèles vont être entraînés et testés.

Étant donné le caractère temporel de la disponibilité des données il serait peu pertinent de choisir aléatoirement les ensembles de test et d'entraînement. Il sera préféré une approche de back testing, l'entraînement se fera donc uniquement sur des données antérieures à celles du test. Concrètement, on commence par fixer une date de test sur laquelle on va extraire les variables prédictives et la variable cible au moyen de la procédure décrite ci-dessus. Ensuite, on choisit plusieurs dates pour l'entraînement qui sont antérieures d'au moins deux semaines à la date de test sur lesquelles on réalise la même extraction de variables.

Comme évoqué en fin de section 2.2.2 le choix des dates doit se faire de manière à être cohérent entre ensembles de test et d'entraînement. Cette cohérence peut être assurée en choisissant des dates d'entraînement proches de la date de test pour être le moins possible exposé à des changements de pratiques dans la gestion des dossiers. On peut également annuler des effets de saisonnalité hebdomadaire en choisissant des dates au même jour de la semaine. Après différents choix testés (notamment des choix de dates aléatoires) la solution finalement retenue sera la suivante.

Définition 2.9 (Ensembles de test et d'entraînement)**.** Si on note X_t et Y_t les tableaux de données des variables prédictives et variable à prédire construites selon la procédure décrite précédemment et t_{test} la date de test, on définit les variables de test et d'entraînement par

$$
\begin{aligned}
X_{test} &= X_{t_{test}}, \\
Y_{test} &= Y_{t_{test}}, \\
X_{entraînement} &= CONCAT\left((X_{t_{test}-7k})_{k\in\{2,3,\ldots,26\}}\right), \\
Y_{entraînement} &= CONCAT\left((Y_{t_{test}-7k})_{k\in\{2,3,\ldots,26\}}\right).
\end{aligned}
$$

Avec $CONCAT$ la fonction qui concatène verticalement des tableaux de données.

Le choix d'utiliser 25 dates revient à s'accorder 6 mois d'entraînement et si on note n_t le nombre de dossiers ouverts en t l'ensemble de test comporte $n_{t_{test}}$ individus (lignes) et l'ensemble d'entraînement en comporte $\sum_{k=2}^{26} n_{t_{test}-7k}$. Dans toute la suite on fixe la date de test au 10/11/2022 ce qui donne 41 388 individus (dossiers en cours) pour le test et 1 038 385 pour l'entraînement.

2.3 Quelques statistiques descriptives

On présente ici quelques statistiques descriptives des variables ainsi que leur comportement en fonction de la présence ou non d'un appel à deux semaines.

2.3.1 Représentations univariées

Variable à prédire

La première statistique d'importance est la proportion de dossiers ayant un appel dans les deux semaines. On a

$$\mathbb{E}(Y_{test}) = 0,077 \text{ et } \mathbb{E}(Y_{entraînement}) = 0,070.$$

On est donc en présence de données assez déséquilibrées ce qui justifie bien de modéliser une probabilité plutôt qu'une classe directement, le modèle constant prédisant la non-présence d'appel étant correct dans la plupart des cas mais peu utile. L'enjeu est donc bien plus porté sur la capacité à identifier des dossiers à risque élevé d'appel plutôt que des dossiers à faible risque d'appel. Ce déséquilibre montre également la difficulté de mise en place de mesures ayant pour but de limiter le nombre d'appels car des mesures généralisées sur l'ensemble du portefeuille seraient au mieux inutiles et au pire pourraient susciter des effets non désirés et augmenter le nombre d'appels.

Principales caractéristiques

La typologie de sinistre, le mode de gestion et le canal de déclaration sont des variables considérées constantes au cours du temps et ne dépendent donc pas de la date où on scanne le portefeuille et on peut se référer aux figures 1.5, 1.6 et 1.7 pour visualiser la répartition de ces variables sur l'ensemble du portefeuille. En revanche, bien que ces variables soient indépendantes de la date de scan l'observation d'un dossier ouvert ne l'est pas. En effet, si les dossiers possédant une certaine caractéristique ont tendance à se clôturer plus rapidement que la moyenne ils seront alors moins représentés au moment d'un scan alors des dossiers ayant une caractéristique selon laquelle le temps de gestion est long seront surreprésentés par la présence à date de scan de dossiers ouverts depuis longtemps.

Pour illustrer ce phénomène on donne la répartition de la typologie de sinistre, du mode de gestion et du canal de déclaration aux figures 2.2, 2.3 et 2.4 sur l'ensemble des données d'entraînement et de test. Certains regroupements ont été effectués de la même façon qu'aux figures 1.5, 1.6 et 1.7. La répartition

du canal de déclaration semble donc être assez similaire à ce qu'elle est sur l'ensemble du portefeuille de l'étude. Cette remarque vaut aussi pour la typologie de sinistre bien que les événements climatiques semblent surreprésentés. En revanche au niveau du mode de gestion on peut constater d'importants changements, les dossiers sans suite sont sous représentés presque de moitié alors que les modes de gestion impliquant de l'expertise ou une réparation en nature sont largement surreprésentés et on remarque également que les dossiers faisant intervenir une indemnisation en gré-à-gré sont à l'image des dossiers sans suite sous-représentés.

Pour illustrer le lien de ces constats avec le délai de gestion jusqu'à la clôture on donne sur la figure 2.5 la distribution de ce délai en fonction du mode de gestion et on a ajouté la distribution concernant les événements climatiques. On peut donc constater que les dossiers des modes de gestion ne faisant pas intervenir d'autre professionnel que le gestionnaire se clôturent plus rapidement et explique donc leur sous-représentation et vice-versa pour les autres modes de gestion. En revanche, les dossiers d'événements climatiques semblent avoir un délai de gestion avant clôture distribué similairement à l'ensemble du portefeuille de l'étude, la surreprésentation de ces dossiers au moment des scans d'entraînement et de test vient donc plus probablement de la variabilité de la survenance des événements climatiques avec une forte activité climatique sur les trois premiers trimestres de 2022.

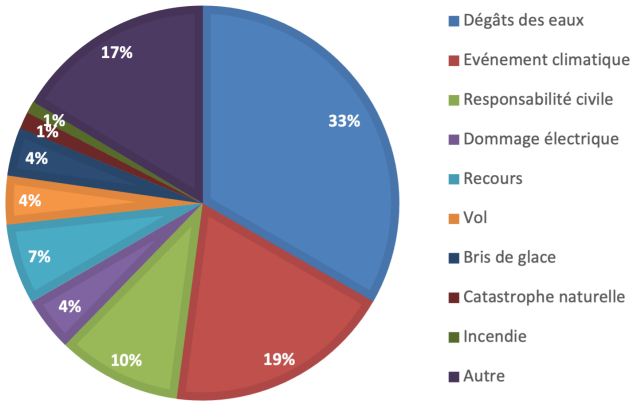


FIGURE 2.2 : Proportions des différentes typologies de sinistre au moment du scan

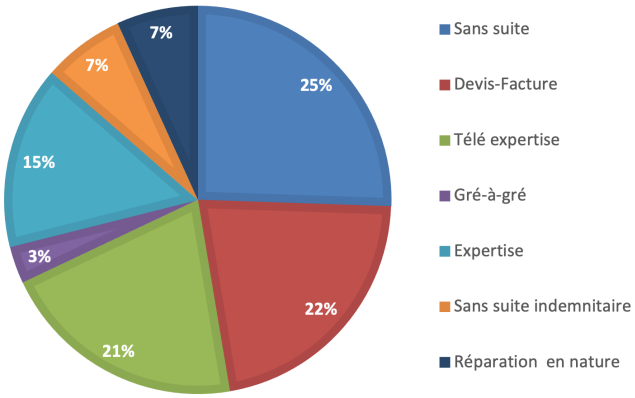


FIGURE 2.3 : Proportions des différents modes de gestion au moment du scan

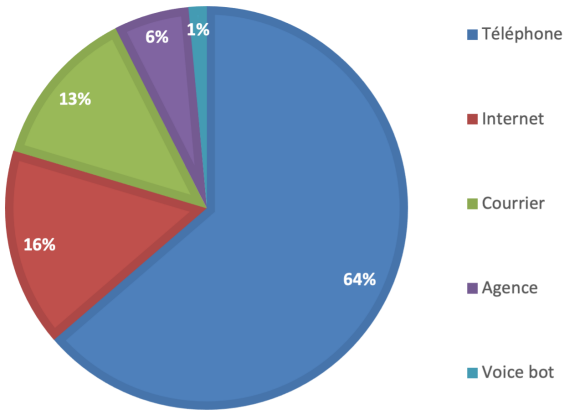


FIGURE 2.4 : Proportions des différents modes de contact à la déclaration au moment du scan

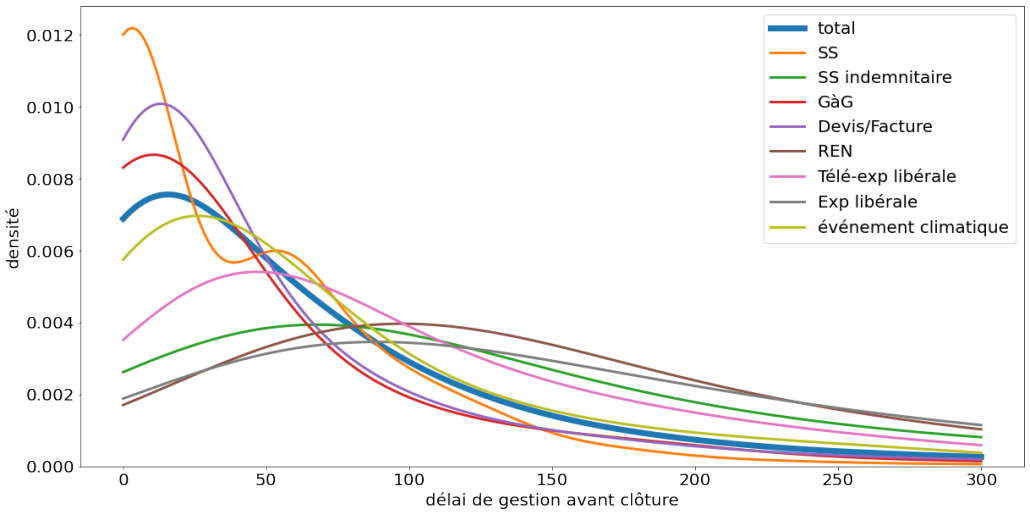


FIGURE 2.5 : Distribution du délais de gestion par mode de gestion et pour les dossiers d'événements climatiques

Temps écoulé depuis l'ouverture

En lien avec les considérations précédentes la "durée de vie" écoulée du dossier au moment du scan a son importance d'une part car elle fait partie de nos variables prédictives et d'autre part car elle permet de juger la nature des dossiers que l'on regarde, à savoir si le dossier en est plutôt dans ses premiers jours ou semaines de gestion ou s'il est plutôt dans une situation où la gestion est rentrée dans une phase plus longue sans avoir pu clôturer rapidement le dossier. La distribution de cette variable est donnée à la figure 2.6 sur laquelle on peut voir que l'on a affaire à des dossiers d'âges très variés bien qu'assez logiquement les plus récents soient les plus représentés. En moyenne les dossiers sont ouverts depuis 127 jours (75 en médiane) pour un écart-type de 135 jours.

On est donc en présence d'une certaine hétérogénéité quant à durée passée en gestion des dossiers et ce point doit être gardé à l'esprit pour ne pas tirer de conclusions trop hâtives sur certaines statistiques

descriptives ou résultats de modélisation notamment en ce qui concerne les modes de gestion. Ces derniers donnant lieu à des durées de gestion assez variables on ne pourrait donc pas juger la capacité globale d'un mode de gestion à générer des appels de gestion en se basant sur des données issues d'un scan. En effet, un dossier gérer en Devis-Facture ayant généralement une clôture plus rapide, s'il est ouvert à date c'est soit qu'il en est dans ses premiers jours ou semaines de gestion ou soit qu'il est ouvert depuis une durée anormalement longue et est donc probablement un dossier complexe, dans les deux cas on peut se dire que la probabilité d'appel devrait en être augmentée.

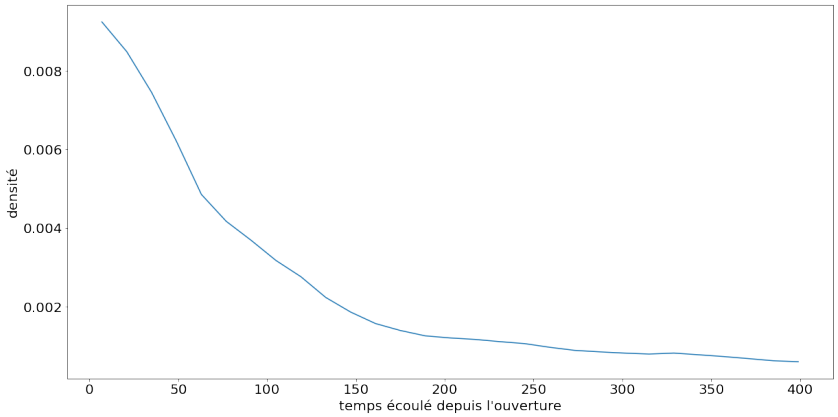


FIGURE 2.6 : Distribution du temps écoulé depuis l'ouverture

Temps écoulé depuis le dernier événement

Le temps écoulé depuis le dernier événement (variable δ dans le scan) permet d'aller observer si le dossier a connu une activité récente ou si ce dernier se situe plus dans une situation "d'attente" du point de vue de la gestion. Cette situation pourrait correspondre à une attente de séchage dans le cadre d'un dégât des eaux ou une période entre la décision du missionnement d'un expert et la remise de son rapport d'expertise par exemple mais peut très bien aussi correspondre à des dossiers dont la gestion est terminée en pratique mais qui n'aurait pas encore été clôturé par le gestionnaire dans le système d'information. La distribution de cette variable est donnée dans la figure 2.7. En moyenne les dossiers sont "inactifs" depuis 57 jours (27 en médiane) pour un écart-type de 78 jours.

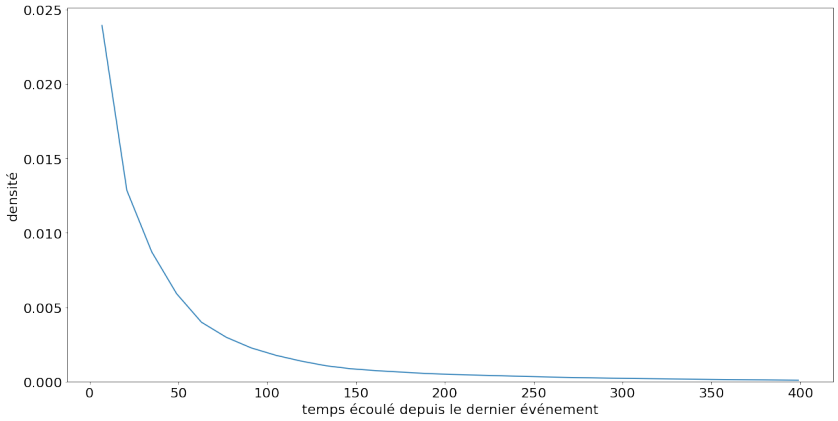


FIGURE 2.7 : Distribution du temps écoulé depuis le dernier événement

Plus de la moitié des dossiers ont donc eu une activité au moins au cours du dernier mois et la plus forte décroissance de la densité comparé à la figure 2.6 montre que les dossiers ouverts depuis longtemps ne sont pas forcément des dossiers inactifs depuis longtemps. On a par exemple 38% des dossiers ouverts depuis plus de 200 jours qui ont eu un événement au cours des 50 derniers jours. On donne de plus la répartition de la nature du dernier événement dans le tableau 2.6.

pli envoyé	pli reçu	appel sortant	appel gestion	ouverture	appel déclaration	mission
41,2%	24,5%	12,6%	11,1%	5,9%	4,2%	0,6%

TABLE 2.6 : Proportion de la nature du dernier événement observé

Intérêt Client

La dernière variable qu'on se propose d'observer dans cette section est la situation du dossier en termes d'intérêt client, sur l'ensemble des scans considérés 67% des dossiers étaient en intérêt client. Il y a donc en moyenne un tiers des dossiers du stock courant qui sont des dossiers pour lesquels l'assuré s'est déjà vu délivrer sa prestation et où le dossier n'est pas clôturé car il reste encore des actes de gestion à réaliser (si on excepte les clôtures tardives sur le système d'information). Il est notamment possible qu'un recours contre une compagnie adverse soit à exercer.

2.3.2 Interactions avec la variable à prédire

Variables quantitatives

Il est intéressant d'aller observer si les variables qu'on a retenues pour la modélisation permettent d'observer une différence de comportement en fonction de la présence ou non d'un appel à deux semaines. Les figures 2.8 et 2.9 donnent les distributions du temps écoulé depuis l'ouverture et de celui écoulé depuis le dernier événement en fonction de la présence ou non d'un appel à deux semaines. Il est possible de voir dans un cas comme dans l'autre que les dossiers appelants sont plus récents que les autres (de manière générale ou en termes d'activité). La différence semble plus marquée pour le temps écoulé depuis le dernier événement et on s'attend donc à ce que cette variable puisse être un bon indicateur dans la modélisation de la probabilité d'appel. D'une manière générale on s'attend donc à ce que ces deux variables quantitatives puissent permettre d'identifier si un dossier semble "actif" et donc propice à l'appel.

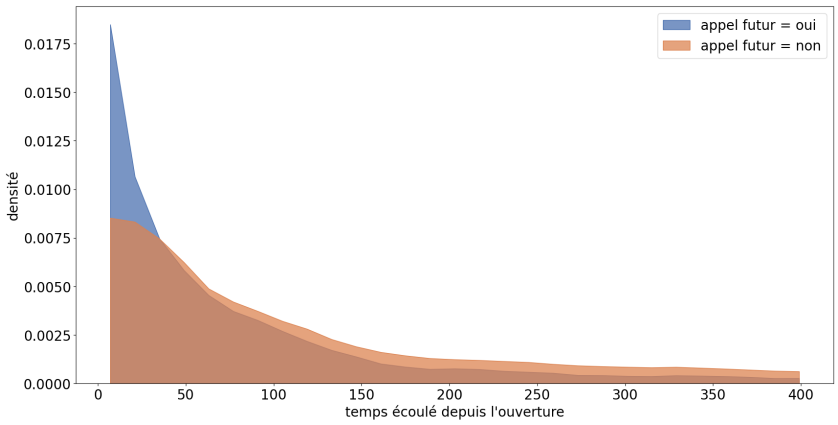


FIGURE 2.8 : Distribution du temps écoulé depuis l'ouverture

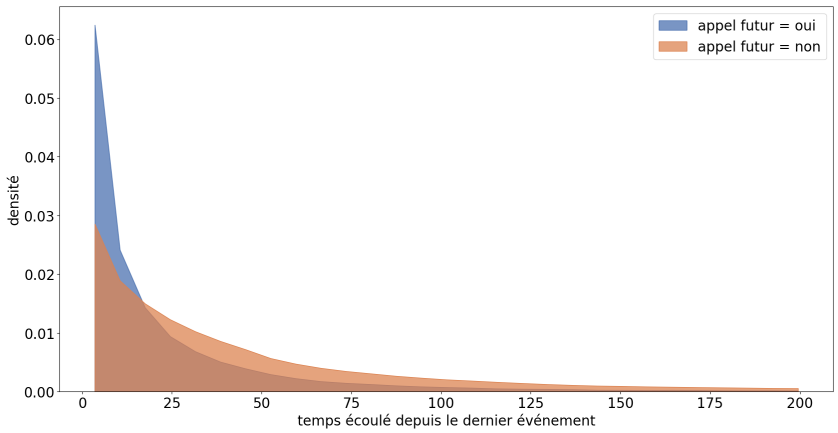
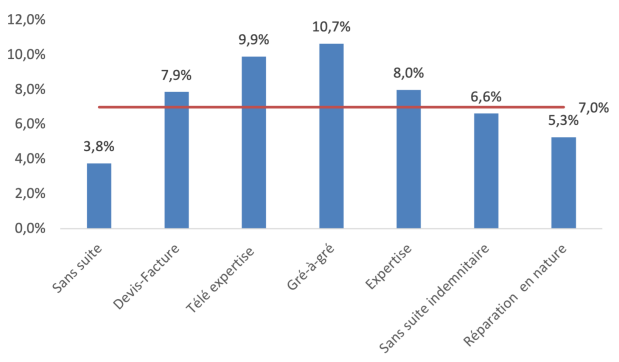


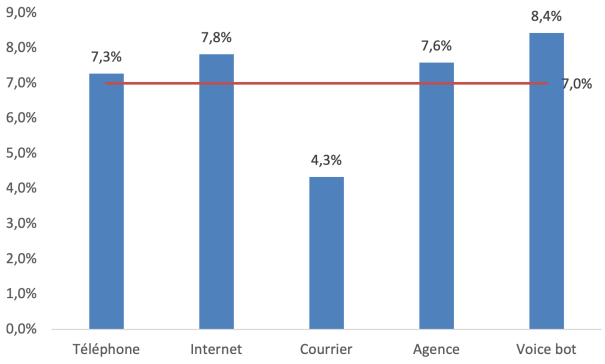
FIGURE 2.9 : Distribution du temps écoulé depuis le dernier événement

Variables qualitatives

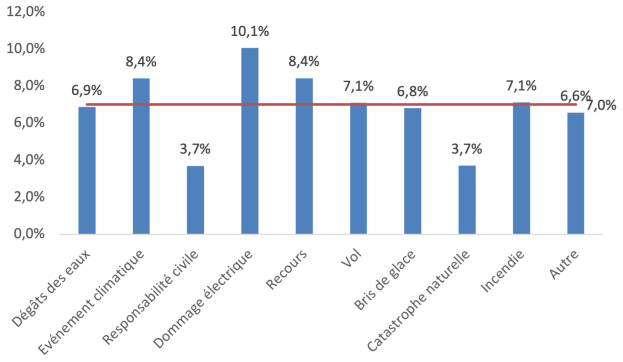
La base de données construite comporte de nombreuses variables qualitatives dont on espère qu'elles puissent permettre de segmenter les dossiers selon des comportements différents. La figure 2.10 donne la proportion des dossiers ayant au moins un appel à deux semaines en fonction du mode de gestion, du mode de contact à la déclaration, de la typologie de sinistre ainsi que de la situation en intérêt client ou non du dossier.



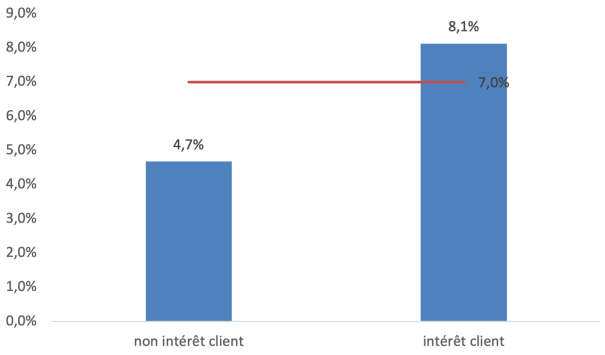
(a) Mode de gestion



(b) Mode de contact à la déclaration



(c) Typologie de sinistre



(d) Intérêt client

FIGURE 2.10 : Taux de dossiers appelants en fonction de certaines variables qualitatives

On peut constater au niveau du mode de gestion que les dossiers gérés en gré-à-gré sont les plus générateurs d'appels, ce constat n'indique pas que ces dossiers comportent plus d'appels au global (voir la figure 1.12 pour ce genre de comparaison). En effet, il faut garder à l'esprit qu'on parle ici de dossiers ouverts et qu'un dossier géré en gré-à-gré qui n'a pas été rapidement clos peut cacher une réalité complexe. Il est également possible de voir qu'assez logiquement les dossiers sans suite sont ceux qui génèrent le moins d'appels. Au niveau du mode de contact à la déclaration le fait assez marquant est que les dossiers ayant été ouverts par courrier sont presque deux fois moins générateurs d'appels que les autres ce qui pourrait traduire une préférence à l'utilisation de la communication écrite. Pour la typologie de sinistre, la responsabilité civile ainsi que les catastrophes naturelles aboutissent au plus bas taux d'appel tandis que les dommages électriques ont un taux sensiblement plus haut. Enfin, assez naturellement les dossiers n'étant plus en situation d'intérêt client occasionnent moins d'appels.

De manière générale le taux d'appel ne semble pas uniforme en fonction des différentes valeurs des variables qualitatives ce qui laisse bon espoir quant à leur utilisation pour modéliser la probabilité d'appel.

2.4 Régression logistique

Dans la mesure où le problème à traiter est un problème de classification binaire il est naturel de s'intéresser à la régression logistique comme premier modèle. Ce dernier nous servira de référence de comparaison par suite. La relative simplicité de ce modèle lui confère une bonne interprétabilité et une bonne explicabilité tout en étant souvent performant ce qui fait qu'il est encore largement utilisé aujourd'hui.

2.4.1 Généralités

La régression linéaire ordinaire est une méthode qui vise à prédire la valeur attendue d'une variable à prédire, une quantité inconnue, en se basant sur une combinaison linéaire de différentes variables prédictives observées. Cette approche s'avère efficace lorsque la variable à prédire peut varier librement sur une échelle continue, ou lorsque sa variation demeure relativement modérée par rapport aux variations des variables prédictives.

Toutefois, cette méthode présente des limites dans certaines situations. Par exemple, lorsque la variable à prédire doit toujours rester positive et que sa variation doit être représentée de manière exponentielle en réponse à des variations des variables prédictives, la régression linéaire classique ne peut pas capturer ce comportement de manière adéquate. De plus, comme dans notre cas, lorsqu'on traite des modèles de probabilité binaire (variable de Bernoulli), où les probabilités sont confinées dans l'intervalle $[0, 1]$, les prédictions linéaires peuvent être inappropriées.

Pour résoudre ces limitations, les modèles linéaires généralisés, Nelder et Wedderburn (1972), offrent une flexibilité accrue en permettant aux variables à prédire de suivre diverses distributions, au-delà de la distribution normale traditionnellement utilisée dans la régression linéaire. De plus, ils permettent l'utilisation de fonctions de liaison non linéaires entre les variables prédictives et la variable à prédire, ce qui signifie que la relation entre ces deux éléments peut être modélisée de manière plus réaliste, sans l'hypothèse restrictive d'une variation linéaire.

Plus formellement un modèle linéaire généralisé garde l'explicatif linaire en supposant

$$g\left(\mathbb{E}(Y_i|X_i)\right) = X_i^T\beta,$$

où X_i^T est un vecteur ligne d'observation des variables prédictives (auxquelles on a précédemment ajouter une variable constante d'intercept) et g est une fonction inversible appelée fonction de lien. On a donc

$$\mathbb{E}(Y_i|X_i) = g^{-1}\left(X_i^T\beta\right),$$

ou encore (en considérant l'ensemble des individus)

$$\mathbb{E}(Y|X) = g^{-1}\left(X^T\beta\right).$$

La fonction de lien dépend donc forcément de la loi à modéliser. Le cas particulier des lois de la famille exponentielle regroupe la plupart des lois usuelles.

Définition 2.10 (Famille exponentielle)**.** On dit qu'un variable aléatoire Y possède une densité de probabilité, par rapport à une mesure dominante ν, notée $f_{\theta,\phi}$ appartenant à la famille exponentielle si $f_{\theta,\phi}$ s'écrit

$$f_{\theta,\phi}(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right),$$

où b et c sont des fonctions connues et dérivables, $\theta \in \mathbb{R}$ est le paramètre naturel et $\phi \in \mathbb{R}_+^*$ est le paramètre de dispersion. On supposera de plus que b est 3 fois dérivable et que b' est inversible

Proposition 2.1. Si Y appartient à la famille exponentielle, en reprenant les notations de la définition on a

$$\mathbb{E}(Y) = b'(\theta) \text{ et } \mathbb{V}(Y) = b''(\theta)\phi.$$

Le choix de la fonction de lien sera souvent celui de la fonction de lien canonique.

Définition 2.11 (fonction de lien canonique)**.** Soit Y une variable aléatoire appartenant à la famille exponentielle, alors la fonction

$$g(\mu) = (b')^{-1}(\mu),$$

est appelée fonction de lien canonique.

Un modèle linéaire généralisé est donc caractérisé par :

- une distribution de probabilité appartenant à la famille exponentielle
- un explicatif linéaire $X^T\beta$
- une fonction de lien g qu'on prendra souvent comme la fonction de lien canonique

Dans notre cas on cherche à prédire une variable de Bernoulli Y qui appartient bien à la famille exponentielle. En effet si $Y \sim \mathcal{B}(p)$ alors

$$\mathbb{P}(Y = y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), y \in \{0,1\},$$

en posant

$$\theta = \ln\left(\frac{p}{1-p}\right),\ b(\theta) = \ln\left(1+e^{\theta}\right),\ \phi = 1 \text{ et } c \equiv 0.$$

La fonction de lien canonique est donc $g(\mu) = (b')^{-1}(\mu) = logit(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$. La régression logistique revient donc à poser le modèle suivant

$$\mathbb{E}(Y|X) = g^{-1}\left(X^T\beta\right) = \frac{1}{1+e^{-X^T\beta}}.$$

La calibration des paramètres (de β) peut se faire par maximum de vraisemblance, pour une loi de la famille exponentielle, en choisissant la fonction de lien canonique, les équations de vraisemblance du modèle se simplifient sous la forme

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)X_{i,j}}{\phi} = 0, \ \ j = 1, ..., p,$$

avec $\mu_i = \mathbb{E}(Y_i \mid X_i)$ et p le nombre de variables de X. Dans le cadre de la régression logistique on a donc

$$\sum_{i=1}^{n} \left(Y_i - \frac{1}{1 + e^{-X_i^T \beta}} \right) X_{i,j} = 0, \ \ j = 1, ..., p. \tag{2.10}$$

Estimer β (les paramètres du modèle) revient donc à résoudre les équations 2.10. On peut donc approcher le maximum de vraisemblance par des méthodes de type Newton ou quasi-Newton par exemple.

2.4.2 Application, performances et résultats

Les résultats présentés ont été obtenus à l'aide de la librairie Python *Scikit-Learn*, Pedregosa et al. (2011), grâce aux ensembles de test et d'entraînement définis précédemment (après normalisation). L'entraînement a pris 51 secondes et la prédiction sur l'ensemble de test moins de 0,1 seconde. La figure 2.11 présente la courbe ROC de l'ensemble de test et permet d'obtenir une AUC de 0,750. Le modèle de régression logistique permet de plus de détecter 198 appels à capacité donnée de 500.

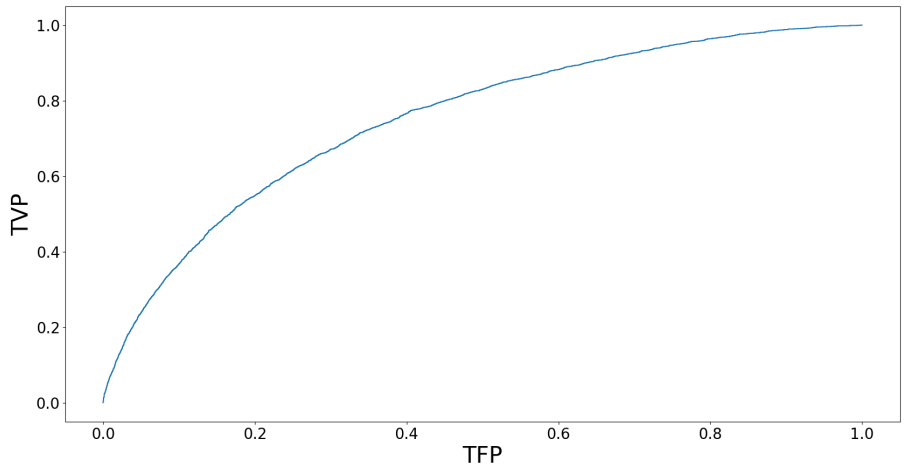


Figure 2.11 : Courbe ROC du modèle de régression logistique

Les performances de ce modèle peuvent donc être considérées comme correctes surtout en intégrant le fait que, comme évoqué en section 1.3.4, les données sur les appels sont incomplètes et il est donc impossible de s'approcher de près du modèle parfait. En effet, il est possible que le modèle identifie un dossier comme ayant une forte probabilité d'appel, qu'un appel intervienne bien en pratique mais que celui-ci n'apparaisse pas dans les données et que donc cette prédiction correcte vienne pénaliser les performances observées.

Si les performances sont correctes, il est intéressant de voir si le modèle ne souffre pas de surapprentissage. Un surapprentissage se caractériserait par exemple par un écart de performance entre

ensemble de test et d'entraînement, cependant sur l'ensemble d'entraînement le modèle obtient une AUC de 0,751 soit sensiblement le même que sur l'ensemble de test. Il est également possible de s'assurer de la cohérence des probabilités d'appel prédites en les regroupant. En effet, si l'on classe les dossiers de l'ensemble de test par leur probabilité d'appel prédite par le modèle et qu'on les segmente en 100 (par exemple) groupes de même volume, on devrait observer une similarité entre la prédiction moyenne et la proportion réelle d'appel si le modèle donne des prédictions cohérentes. La figure 2.12 montre bien cette cohérence.

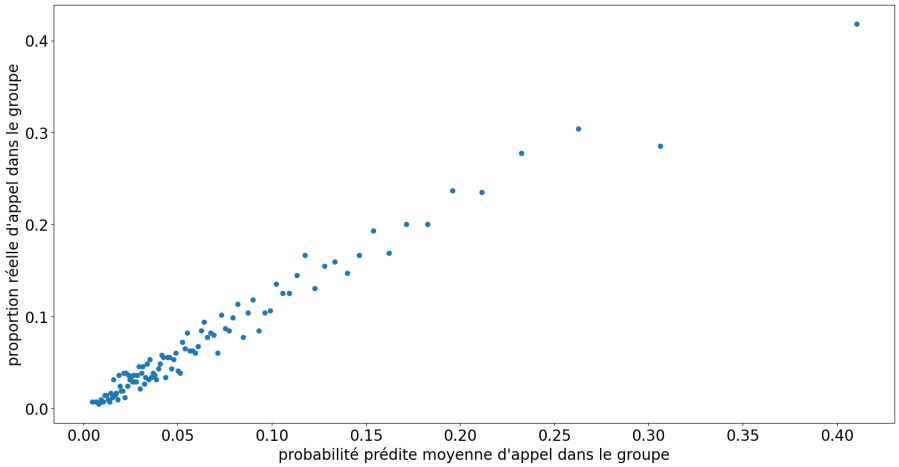


FIGURE 2.12 : Cohérence du modèle de régression logistique

2.4.3 Sélection/importance des variables

Les coefficients estimés et les erreurs d'estimation de la régression logistique sont donnés dans le tableau 2.7 par ordre décroissant de valeur absolue, ils suggèrent qu'un variable a une importance dominante, il s'agit de δ (le temps qu'il s'est écoulé depuis le dernier événement). Cet effet se traduit par une probabilité d'appel prédite d'autant plus grande que le dossier a eu une activité récente. Cet effet est observable sur la figure 2.13 où est tracée la distribution de δ sur des groupes classés par probabilité d'appel prédite. On y voit que les 5% de dossier avec la plus haute prédiction ont majoritairement eu une interaction au cours de la dernière semaine.

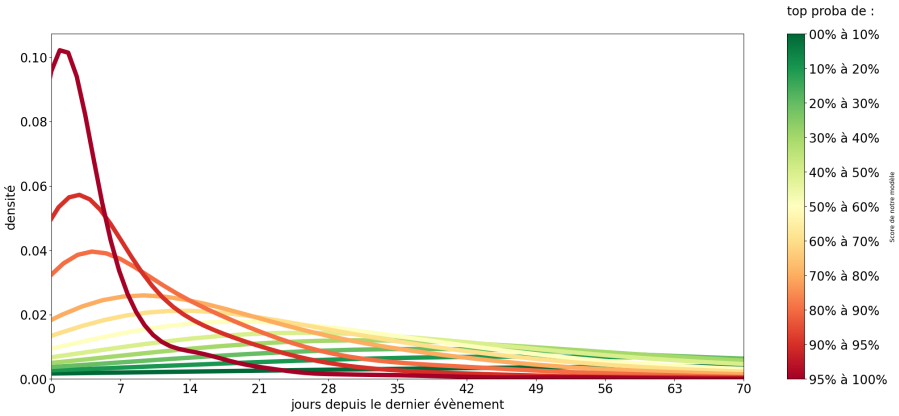


FIGURE 2.13 : Distribution de δ en fonction de la prédiction du modèle de régression logistique

variable	coefficient	S.D.	variable	coefficient	S.D.
δ	-0,5	0,006	e_{-5} : -	0,012	0,012
δ_{ouv}	-0,14	0,005	e_{-2} : mission	-0,012	0,004
e : appel e gestion	0,069	0,022	typo : Autres Dommages	0,011	0,008
MG : SS	-0,068	0,021	typo : Incendie	0,011	0,003
typo : Individuelle	0,05	0,173	e_{-1} : appel sortant	0,011	0,011
e : ouverture	-0,05	0,12	e_{-1} : pli reçu	0,011	0,009
typo : Recours	0,045	0,023	CD : Internet	0,011	0,005
e_{-1} : appel e gestion	0,044	0,026	δ_{-5}	-0,011	0,011
MG : GàG	0,044	0,027	CD : Courrier	-0,011	0,001
δ_{-1}	-0,044	0,011	CD : Chatbot	0,011	0,01
e : mission	-0,043	0,039	e_{-3} : ouverture	-0,01	0,004
MG : Télé-exp libérale	0,041	0,024	e_{-1} : pli envoyé	-0,01	0,008
typo : Responsabilité Civile	-0,04	0,005	e_{-2} : pli reçu	0,01	0,008
e_{-2} : appel e gestion	0,038	0,011	e_{-4} : -	0,01	0,005
CD : Réseaux scx	-0,037	0,038	e : pli envoyé	-0,01	0,006
e_{-3} : appel e gestion	0,037	0,006	MG : SS indemnitaire	0,009	0,005
e_{-1} : appel de déclaration	0,036	0,022	e_{-1} : -	0,009	0,005
e_{-2} : appel de déclaration	0,036	0,009	typo : Evénement Climatique.	0,009	0,005
CD : Sté assistance	0,034	0,027	CD : Visite	0,008	0,001
e : appel de déclaration	0,033	0,017	CD : EDI	-0,008	0,001
e : pli reçu	0,032	0,008	e_{-3} : appel sortant	0,008	0,003
e_{-4} : appel e gestion	0,031	0,027	e_{-3} : pli reçu	0,008	0,004
CD : VoiceBot	0,03	0,005	δ_{-4}	-0,008	0,008
MG : Exp libérale	0,028	0,018	e_{-2} : -	0,007	0,001
e_{-5} : appel e gestion	0,028	0,016	typo : Vol	0,007	0,002
δ_{-3}	-0,026	0,009	typo : Catastrophe Naturelle.	-0,007	0,003
CD : Mobile	0,026	0,004	e_{-2} : appel sortant	0,006	0,006
e_{-3} : appel de déclaration	0,026	0,025	MG : REN	-0,006	0,005
e_{-5} : appel de déclaration	0,026	0,025	typo : Explosion Véhicule	-0,006	0,006
e : appel sortant	0,025	0,016	e_{-4} : pli reçu	0,005	0,003
typo : Autres IRD	0,024	0,003	e_{-2} : pli envoyé	-0,005	0,003
e_{-4} : appel de déclaration	0,023	0,016	e_{-4} : appel sortant	0,005	0,005
MG : Autre	-0,022	0,02	e_{-5} : ouverture	0,005	0,003
typo :	0,022	0,015	e_{-5} : pli envoyé	-0,004	0,001
IC	0,022	0,016	e_{-5} : pli reçu	0,003	0,001
e_{-1} : mission	-0,021	0,01	e_{-4} : pli envoyé	-0,003	0,003
CD : Intranet	0,019	0,006	e_{-3} : -	0,003	0,002
δ_{-2}	-0,018	0,017	e_{-5} : appel sortant	-0,002	0,001
CD : Téléphone	0,018	0,009	e_{-3} : mission	-0,002	0,002
typo : Dégâts des eaux	0,018	0,006	typo : Bris de glace	0,002	0,002
typo : FMP	-0,017	0,003	e_{-4} : ouverture	-0,001	0,001
typo : Dommage électrique	0,017	0,001	e_{-4} : mission	-0,001	0,001
e_{-2} : ouverture	-0,013	0,007	e_{-3} : pli envoyé	-0,001	0,001
e_{-1} : ouverture	-0,013	0,005	e_{-5} : mission	0,001	0,001
MG : Devis/Facture	0,013	0,003			

TABLE 2.7 : Coefficients de la régression logistique par ordre décroissant de valeur absolue

En analysant les autres coefficients et les résultats du modèle il est également possible de constater que la durée depuis l'ouverture vient diminuer la probabilité d'appel (plus un dossier est ouvert récemment plus sa probabilité d'appel est grande) ou encore par exemple que les dossiers gérés en expertise ou en télé-expertise voient leurs prédictions augmenter. Autre fait intéressant, le poids accordé à la présence d'un appel de gestion pour dernier événement qui laisse suggérer une certaine forme de répétition du comportement.

Il est possible de considérer la valeur absolue des coefficients comme une mesure de l'importance des variables. La grande disparité de ces coefficients pourrait laisser imaginer une sélection de variables par pénalisation de la norme 1 des coefficients (Lasso) ou par des méthodes pas à pas. Cependant, dans la mesure où une sélection de variables n'a pas amené à une amélioration des performances sur l'ensemble de test et que l'on souhaite pouvoir analyser l'impact de chacune des variables il a été choisi de garder l'ensemble de ces dernières.

2.5 Arbre de classification

La contrainte de l'explicatif linéaire imposé par le modèle de régression logistique limite sa capacité à capter tous les effets des variables explicatives notamment les interactions et les impacts non monotones. Cette contrainte peut être relâchée en utilisant des arbres de décision qui sont des modèles non-paramétriques simples de conception.

2.5.1 Généralités

On s'intéresse ici aux arbres de décisions et à leur construction tels que définis dans CART (Classification And Regression Trees), Breiman et al. (1984), bien qu'ils existent d'autres méthodes (ID3, C4.5 ou C5.0). Les arbres de décision sont une méthode fondamentale dans l'apprentissage automatique. Ils sont utilisés pour déterminer la variable à prédire en appliquant une séquence de tests binaires sur les variables prédictives comme l'illustre la figure 2.14. Ces tests sont organisés sous forme d'une structure hiérarchique en forme d'arbre où chaque nœud représente un test sur une variable. Les résultats de ces tests guident le processus de sélection des chemins à travers l'arbre, et cela de manière récursive jusqu'à ce qu'une feuille de l'arbre soit atteinte. Chaque feuille est associée à un groupe d'individus qui doivent donc avoir des valeurs cohérentes de la variable à prédire.

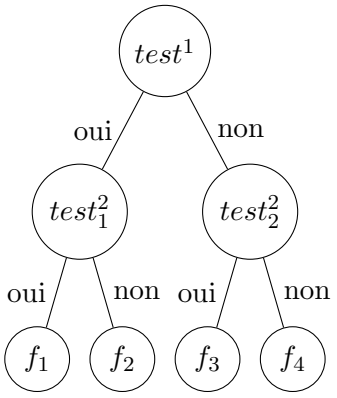


Figure 2.14 : Un exemple d'arbre de décision

La construction de l'arbre se fait sous la méthode suivante, à chaque nœud de l'arbre une variable prédictive et un test sur cette dernière sont choisis de manière à séparer le mieux possible les deux

groupes engendrés selon la valeur de la variable à prédire. Dans la cas d'un arbre de classification pour une variable à prédire catégorielle Y ayant $1, 2, ..., K$ pour modalités et avec X les variables prédictives la procédure de construction de l'arbre est la suivante.

Un nœud qui représente une région R pour les variables prédictives est découpé en deux sous-régions R_1 et R_2 définies grâce à la variable j et au seuil s par

$$R_1(j, s) = \{i \mid X_{i,j} \leq s\} \text{ et } R_2(j, s) = \{i \mid X_{i,j} > s\}.$$

La variable j et le seuil s sont choisis de sorte à minimiser l'impureté engendrée par la coupure définie par

$$\frac{N_1}{N} H(R_1) + \frac{N_2}{N} H(R_2), \tag{2.11}$$

avec N, N_1 et N_2 le nombre d'individus présents dans les régions R, R_1 et R_2 et H une mesure d'impureté d'une région. Dans la suite l'indice de Gini sera utilisé comme mesure d'impureté.

Définition 2.12. Soit R_s une région comportant N_s individus, on note p_k^s la proportion d'individus appartenant à la classe k au sein de R_s, soit

$$p_k^s = \frac{1}{N_s} \sum_{X_i \in R_s} \mathbb{1}_{Y_i = k} \; k = 1, ..., K.$$

On définit alors l'indice de Gini de la région R_s, noté $H_g(R_s)$ par

$$H_g(R_s) = \sum_{k=1}^{K} p_k^s (1 - p_k^s).$$

Remarque 2.4. Pour la prédiction l'arbre de classification renvoie les proportions de chaque modalité de la feuille atteinte lors du parcours de l'arbre selon les variables prédictives.

Cette procédure permet donc de construire un arbre de classification qui sépare progressivement l'espace des variables prédictives en des régions de plus en plus petites. En l'absence de critère supplémentaire, la procédure se poursuit jusqu'à obtenir des régions ne contenant que des individus ayant la même modalité pour la variable à prédire (sauf si certains individus ont exactement les mêmes valeurs de variables prédictives sans avoir la même modalité de variable à prédire).

Construire un arbre sans contrainte sur un ensemble d'entraînement va donc conduire à un surapprentissage et des performances médiocres sur un ensemble de test. En effet, l'arbre va réussir à identifier les individus de l'ensemble d'entraînement plutôt que des groupes cohérents. Pour se prémunir de cet effet plusieurs solutions ont été développées, on peut par exemple limiter la profondeur de l'arbre ou ne subdiviser un nœud que s'il comporte un certain nombre d'individus ou si cela entraîne au moins une certaine réduction de l'impureté soit $H(R) - \frac{N_1}{N} H(R_1) + \frac{N_2}{N} H(R_2) \geq r$, avec r la réduction minimale. On peut également choisir de construire un arbre de grande taille et en extraire un arbre plus simple en l'élaguant.

Élaguer un arbre consiste à réaliser un arbitrage entre performance (sur l'ensemble d'entraînement) et complexité. On définit pour cela le coût d'un arbre comme suit.

Définition 2.13 (Coût d'un arbre)**.** Soient T un arbre de classification et $\alpha \geq 0$ un paramètre de pénalisation de la complexité de l'arbre, on définit le coût de l'arbre, noté $C_\alpha(T)$ par

$$C_\alpha(T) = R(T) + \alpha |T|, \tag{2.12}$$

avec $\alpha|T|$ le nombre de nœuds terminaux (feuilles) dans T et R une mesure d'erreur de l'arbre.

$R(T)$ est habituellement choisi comme la proportion d'individus (dans l'ensemble d'entraînement) qui sont mal classé par T lorsque la classification se fait par vote majoritaire. Une solution alternative (utilisée dans *Scikit-Learn*) est de prendre $R(T)$ comme la moyenne pondérée de la mesure d'impureté des nœuds terminaux. Élaguer un arbre T revient donc à trouver le sous-arbre qui minimise l'équation 2.12, un sous-arbre étant défini en supprimant des nœuds non-terminaux (et leurs descendants). Pour $\alpha \geq 0$, on peut montrer qu'il existe un unique plus petit sous-arbre T_α qui minimise l'équation 2.12.

La recherche de T_α peut se faire selon la procédure suivante. On définit le coût d'un nœud t par $C_\alpha(t) = R(t) + \alpha$, en considérant le nœud comme un arbre (ayant uniquement une racine et une feuille confondues en t). On définit également T_t comme l'arbre extrait de T en considérant t comme racine. Le choix d'élaguer ou non le nœud t se fait donc en comparant $C_\alpha(T_t)$ et $C_\alpha(t)$ or cette comparaison dépend de α et on pose donc α_t tel que $C_{\alpha_t}(T_t) = C_{\alpha_t}(t)$ soit

$$\alpha_t = \frac{R(t) - R(T_t)}{|T_t| - 1}.$$

Le nœud t ayant le α_t minimal est donc le nœud le plus faible et est élagué. On répète l'opération jusqu'à ce qu'il n'y ait plus de nœud ayant un $\alpha_t > \alpha$.

Le paramètre α peut être calibré par validation croisée. Par la suite il a été choisi de préférer un critère de nombre minimal d'individus pour diviser un nœud (5000 en pratique) car cela a conduit à de meilleures performances. Cela peut s'expliquer par la quantité de données d'entraînement car une règle de classification qui n'apporte que peu de réduction de l'erreur pour une grande profondeur d'arbre peut tout de même être robuste si elle se base sur une large population d'observations.

2.5.2 Application, performances et résultats

Les résultats présentés ont été obtenus à l'aide de de la librairie Python *Scikit-Learn* grâce aux ensembles de test et d'entraînement définis précédemment. L'entraînement a pris 39 secondes et la prédiction sur l'ensemble de test moins de 0,1 seconde. La figure 2.15 présente la courbe ROC de l'ensemble de test et permet d'obtenir une AUC de 0,745. Le modèle d'arbre de classification permet de plus de détecter 171 appels à capacité donnée de 500.

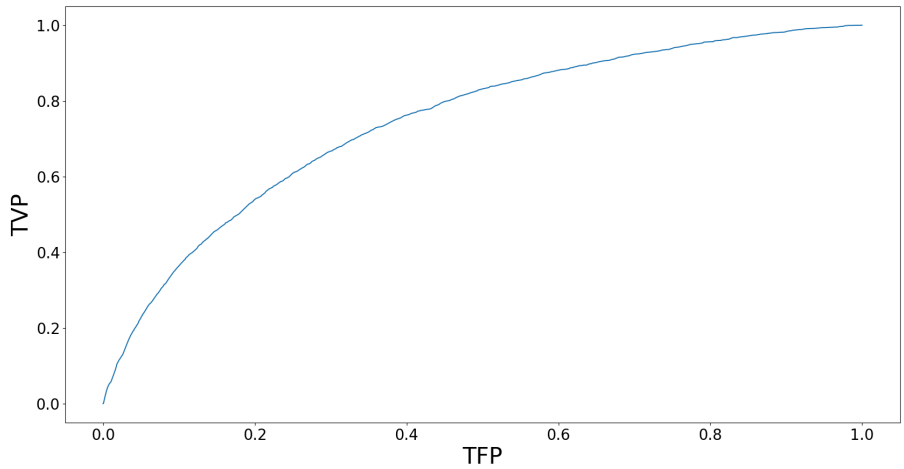


FIGURE 2.15 : Courbe ROC du modèle d'arbre de classification

Les performances du modèle d'arbre de classification sont donc correctes mais légèrement inférieures

à celles de la régression logistique. De la même façon que pour la régression on peut s'assurer de l'absence de surapprentissage par des performances similaires sur les ensembles d'entraînement et de test (AUC de 0,756 sur l'ensemble d'entraînement). On peut constater une cohérence des valeurs prédites sur la figure 2.16 construite de la même façon que précédemment (figure 2.12).

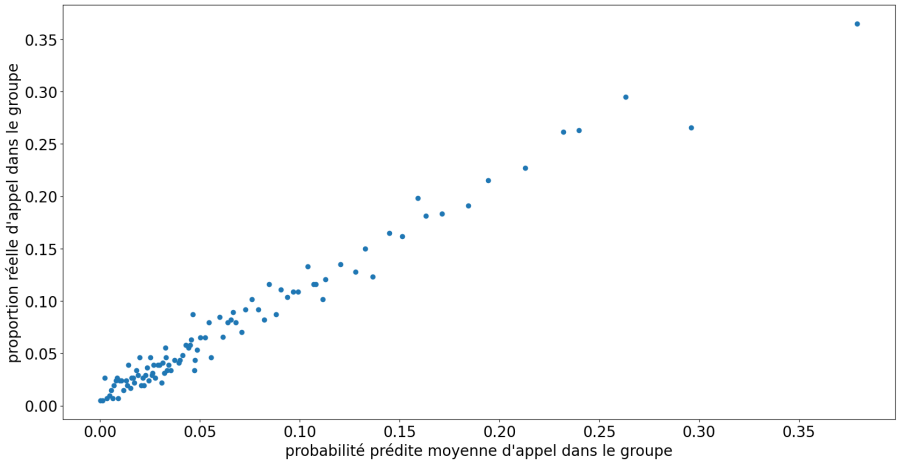
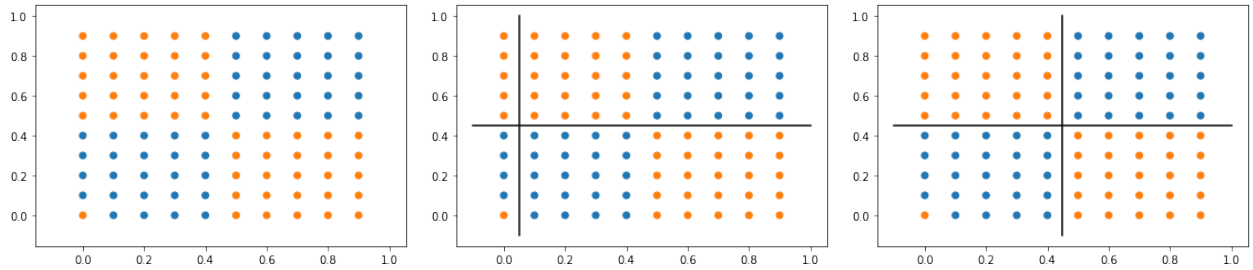


FIGURE 2.16 : Cohérence du modèle d'arbre de classification

Un arbre de classification doit permettre de capter des relations plus complexes que la régression logistique, la comparaison des performances pourrait laisser penser l'inverse. Il est important de garder à l'esprit que la méthode de construction de l'arbre ne garantit en rien de trouver le meilleur arbre pour une complexité donnée. En d'autres termes, si par exemple on fixe une profondeur maximale il est parfois possible de trouver des arbres de cette profondeur plus performants que celui donné par la méthode CART. Cela est largement visible sur la figure 2.17 qui illustre la difficulté de la procédure de la construction de l'arbre à bien identifier les effets croisés.



(a) Des données à classifier (b) CART (c) Un arbre plus performant

FIGURE 2.17 : À gauche des données à classifier, au centre les régions déterminées par la méthode CART pour une profondeur d'arbre de 2 et à droite les régions déterminées par un autre arbre de classification de profondeur 2.

2.5.3 Sélection/importance des variables

Pour un arbre de classification il est possible de définir l'importance d'une variable comme la part de réduction de l'impureté qu'elle permet via tous les nœuds où elle intervient. Cette réduction étant définie comme

$$\frac{N}{\mathcal{N}}\left(H(R) - \frac{N_1}{N}H(R_1) - \frac{N_2}{N}H(R_2)\right),$$

avec les mêmes notations que pour l'équation (2.11) et $\mathcal{N}$ le nombre total d'individu.

L'importance de chaque variable calculée selon cette méthode est donnée au tableau 2.8, il est à nouveau remarquable d'observer la prépondérance de la variable δ (le temps qu'il s'est écoulé depuis le dernier événement) qui concentre 41,2% de la réduction de l'impureté. Si l'on constate à nouveau que le temps écoulé depuis l'ouverture est également important, avec 9,2% de la réduction de l'impureté, il est intéressant de constater que la présence d'un appel de gestion pour dernier événement connu est la deuxième variable réduisant le plus l'impureté avec 11,6%. Cette observation laisse entendre que la présence récente d'un appel de gestion est un bon indicateur de l'arrivée futur d'un appel de gestion.

Il n'est pas nécessaire de procéder à une sélection de variable à proprement parler étant donné que le modèle sélectionne lui-même les variables qu'il va utiliser pour une complexité de modèle imposée. Il est possible d'effectuer ce constat en observant que certaines variables ont une importance laissée nulle par le modèle.

Afin de visualiser l'utilisation des variables faite par l'arbre de classification pour construire des régions ayant des comportements les plus différenciés possible, on donne le début de l'arbre construit à la figure 2.18 en se limitant à un profondeur de 4. Chaque nœud indique le test binaire réalisé ainsi que la proportion des individus arrivant à ce nœud et leur taux de dossiers appelants (à droite dans value). En réalisant trois tests la région identifiée par le modèle comme étant la plus génératrice d'appels est celle des 2,8% de dossiers ayant eu leur dernier événement il y a moins de 9,5 jours pour lesquels cet événement était un appel de gestion et qui ne sont pas des dossiers sans suite. Ces dossiers ont une probabilité d'appel de 0,257 soit près de 4 fois supérieure au taux moyen.

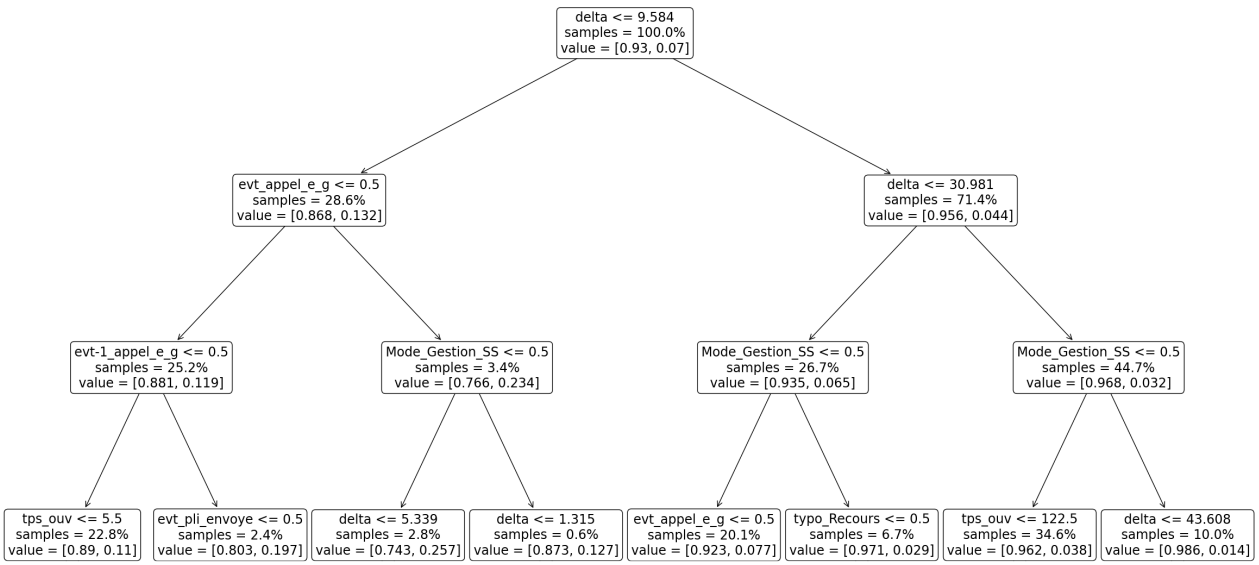


FIGURE 2.18 : Arbre de classification construit tronqué à une profondeur de 4

variable	importance	variable	importance
δ	0,412	typo : Incendie	0,000
e : appel de gestion	0,116	typo : Dégâts des eaux	0,000
δ_{ouv}	0,092	e_{-5} : appel de déclaration	0,000
MG : SS	0,077	e_{-5} : pli envoyé	0,000
e_{-1} : appel de gestion	0,050	e_{-1} : pli reçu	0,000
e_{-2} : appel de gestion	0,029	MG : Autre	0,000
e : pli envoyé	0,029	typo : Individuelle	0,000
typo : Responsabilité Civile	0,022	e_{-2} : ouverture	0,000
e_{-3} : appel de gestion	0,020	e_{-5} : pli reçu	0,000
δ_{-1}	0,019	typo : Catastrophe Naturelle.	0,000
e : pli reçu	0,017	e_{-1} : ouverture	0,000
e : appel de déclaration	0,011	e_{-4} : pli envoyé	0,000
e_{-4} : appel de gestion	0,009	e_{-2} : pli reçu	0,000
δ_{-2}	0,009	CD : Sté asst	0,000
CD : Courrier	0,008	e_{-3} : pli envoyé	0,000
e_{-5} : appel de gestion	0,008	e_{-5} : appel sortant	0,000
e_{-1} : appel de déclaration	0,008	e_{-2} : pli envoyé	0,000
δ_{-3}	0,006	e_{-5} : mission	0,000
typo : Recours	0,005	e : mission	0,000
IC	0,004	e_{-5} : ouverture	0,000
δ_{-5}	0,004	e_{-3} : mission	0,000
e_{-1} : pli envoyé	0,004	typo : Autres Dommages	0,000
δ_{-4}	0,003	e_{-4} : -	0,000
MG : REN	0,003	typo : Vol	0,000
e_{-2} : appel de déclaration	0,003	CD : EDI	0,000
CD : Téléphone	0,003	e_{-3} : appel sortant	0,000
MG : GàG	0,002	CD : Visite	0,000
e_{-3} : appel de déclaration	0,002	e_{-4} : appel sortant	0,000
MG : Télé-exp libérale	0,002	e_{-2} : appel sortant	0,000
e : appel sortant	0,002	CD : Mobile	0,000
CD : VoiceBot	0,001	typo : Explosion Véhicule	0,000
typo :	0,001	e_{-4} : mission	0,000
CD : Chatbot	0,001	e_{-3} : pli reçu	0,000
e_{-2} : mission	0,001	e_{-3} : ouverture	0,000
e_{-2} : -	0,001	e_{-4} : ouverture	0,000
MG : Devis/Facture	0,001	typo : FMP	0,000
MG : Exp libérale	0,001	e_{-4} : pli reçu	0,000
e_{-1} : mission	0,001	e_{-5} : -	0,000
typo : Evénement Climatique.	0,001	e_{-3} : -	0,000
e_{-1} : appel sortant	0,001	e : ouverture	0,000
typo : Dommage électrique	0,001	MG : SS indemnitaire	0,000
typo : Autres IRD	0,001	CD : Internet	0,000
e_{-4} : appel de déclaration	0,001	CD : Intranet	0,000
CD : Réseaux scx	0,001	e_{-1} : -	0,000
typo : Bris de glace	0,000		

TABLE 2.8 : Importance des variables dans l'arbre de classification par ordre décroissant

2.6 D'autres outils

Les deux modèles précédents ont donc permis d'aboutir à des performances correctes en captant bien certains des effets observés avec les statistiques descriptives. Afin d'améliorer les performances de la modélisation, trois approches plus complexes (et donc moins explicables et interprétables) vont être étudiées dont deux réutilisent le principe d'arbre de classification. L'objectif ici sera donc principalement d'améliorer les performances de modélisation dans l'optique de disposer de l'estimateur le plus fiable dans le cas d'une utilisation pratique de la prédiction.

2.6.1 Quelques outils plus puissants

Forêt aléatoire

Une façon d'utiliser les arbres de classification pour construire des modèles plus performants est l'utilisation de forêts aléatoires formalisées par Breiman (2001). Un arbre de classification réalise un compromis entre biais et variance, plus un arbre sera complexe moins il sera biaisé (sur l'ensemble d'entraînement) mais plus ses prédictions seront variables (et donc moins le modèle sera généralisable). L'idée des forêts aléatoire est donc de réaliser la moyenne de plusieurs arbres dans le but de faire diminuer la variance de la prédiction. Pour que ce principe fonctionne il est nécessaire de pouvoir disposer de plusieurs arbres non corrélés. Pour créer ces différents estimateurs (arbres) une forêt aléatoire utilise deux principes.

- Chaque arbre est construit sur un ensemble "bootstrap" de l'ensemble d'entraînement (i.e. un tirage aléatoire avec remise).
- À chaque nœud, la recherche de la division optimale se fait sur sous-ensemble aléatoire de k variables parmi les p variables prédictives. En général (dans le cas d'arbe de classification), on choisit $k = \sqrt{p}$.

La construction des arbres constituant la forêt aléatoire peut aussi se faire avec des contraintes supplémentaires (profondeur maximales, minimum d'individus par nœuds ou encore minimum de réduction de l'impureté) pour éviter un trop grand surapprentissage.

XGBoost

Une autre façon de combiner des arbres de classification dans le but d'obtenir un modèle plus performant est le boosting. Cette pratique consiste à créer un modèle fort en combinant plusieurs modèles faibles (des arbres dans notre cas). Contrairement aux forêts aléatoires qui cherchent à construire des arbres indépendants les méthodes de boosting construisent itérativement des arbres visant à réduire l'erreur des arbres précédents. Le cas particulier du gradient boosting sera la méthode étudiée ici. Le gradient boosting construit itérativement des estimateurs qui viennent prédire les erreurs des estimateurs précédents. Si on cherche à approximer une fonction $f(x)$ on va donc construire itérativement $f_m(x) = f_{m-1}(x) + h_m(x)$ de telle sorte à ce que h_m vienne réduire l'écart entre f et f_{m-1}.

Le terme de gradient boosting vient du fait que si l'on pose une fonction de perte $\mathcal{L}$ différentiable on cherche donc à l'étape m à construire un modèle qui minimise

$$\sum_{i=1}^{n} \mathcal{L}(Y_i, f_{m-1}(X_i)).$$

Si l'on approche bien $r_i^m := -\nabla_{f_{m-1}} \mathcal{L}(Y_i, f_{m-1}(X_i))$ par un modèle (un arbre) $\widehat{r^m}$ alors on peut appliquer une technique de descente de gradient en posant $f_m = f_{m-1} + \gamma \widehat{r^m}$ avec γ le pas utilisé dans la descente de gradient.

XGBoost, Chen et Guestrin (2016), est une version optimisée et open-source du gradient boosting qui intègre notamment des contraintes de pénalisations et dont les performances sur des applications très variées ont conduit à sa popularité.

Réseau de neurones

La dernière solution de modélisation qui sera utilisée sera celle des réseaux de neurones. Plus précisément, on s'intéressera au cas du perceptron multicouche. Un perceptron multicouche se composent des éléments suivants :

- une couche d'entrée qui correspond aux variables prédictives
- une ou plusieurs couches cachées
- une couche de sortie qui dans le cas d'un problème de classification binaire comporte un seul neurone correspondant à la probabilité d'appartenance à la classe positive

Le calcul de la couche de sortie en fonction de la couche d'entrée se fait sous le principe suivant. La valeur d'un neurone se calcule en faisant une somme pondérée des valeurs des neurones (et d'un biais) de la couche précédente sur laquelle on vient appliquer une fonction d'activation (non linéaire). On obtient donc pour le calcul de la valeur du j-ème neurone y_j d'une certaine couche en fonction des n neurones $(x_1, \dots, x_n)$ de la couche précédente, la formule suivante

$$y_j = \sigma\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right),$$

avec σ une fonction d'activation, $(w_{ij})_{i=1,\dots,n}$ les poids des liaisons et b_j le biais. La fonction d'activation peut-être la même sur l'ensemble des neurones du réseau, différente par couche ou encore propre à chaque neurone. On peut remarquer qu'en l'absence de couche cachée (perceptron simple) pour un problème de classification binaire si l'on applique la fonction sigmoïde ($\sigma(x) = \frac{1}{1+e^{-x}}$) au neurone de sortie on se retrouve avec le même modèle que celui défini par la régression logistique.

L'intérêt des couches cachées est donc d'aller capter des effets non-linéaires. Une représentation d'un perceptron multicouche pour une tâche de classification binaire avec deux variables prédictives et deux couches cachées de trois neurones est donnée à la figure 2.19.

L'apprentissage du modèle doit donc permettre d'ajuster le poids de chaque liaison et le biais de chaque neurone des couches cachées et de sortie. Cette calibration peut se faire par un algorithme de rétropropagation du gradient (ou d'autres algorithmes s'en rapprochant) qui se base sur les principes suivants :

- les paramètres du réseau sont initialisés (aléatoirement par exemple)
- les échantillons de l'ensemble d'apprentissage ajuste successivement les poids suivant le protocole suivant
  - calcul de la fonction de perte entre la prédiction pour l'échantillon et la valeur réelle
  - mise à jour des paramètres dans la direction opposée au gradient de la fonction de perte selon un taux d'apprentissage
- utilisation du réseau avec les paramètres calibrés pour réaliser des prédictions

La puissance de la méthode réside en le fait que le calcul du gradient peut se faire en utilisant la formule de dérivation des fonctions composées.

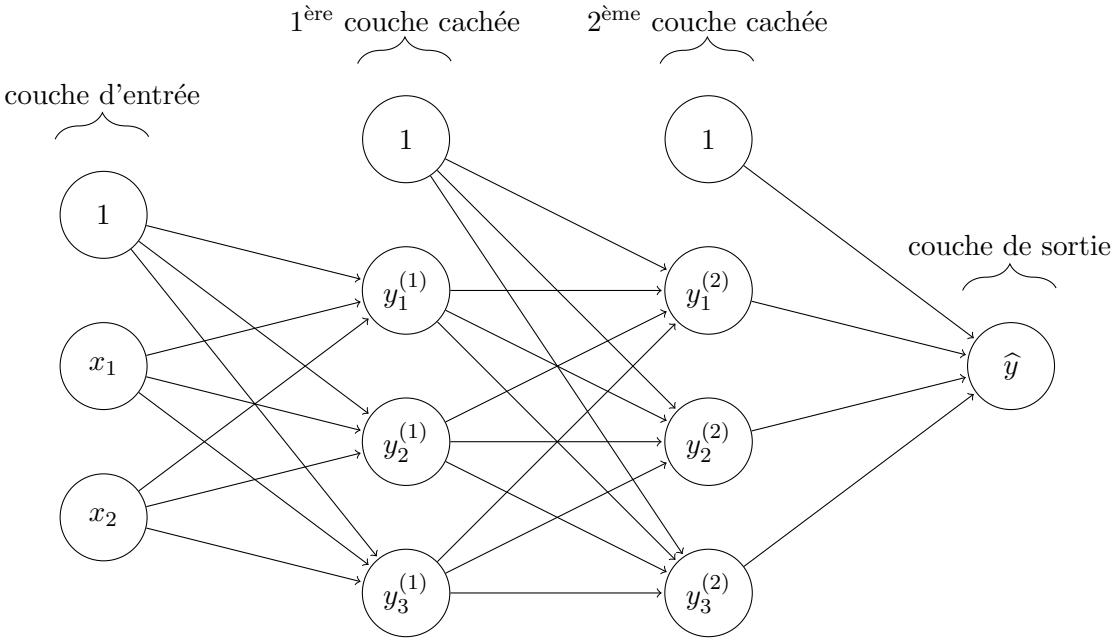


FIGURE 2.19 : Un exemple de structure de perceptron multicouche

2.6.2 Applications, performances et résultats

Les applications présentées ici pour les trois méthodes évoquées ci-dessus ont été réalisées avec la librairie Python *Scikit-Learn* pour la forêt aléatoire et le réseau de neurones et la librairie Python *XGBoost* pour la méthode éponyme. Les courbes ROC ainsi que les graphiques de cohérence des modèles sont donnés aux figures 2.20, 2.21 et 2.22 et les performances sont résumées dans le tableau 2.9.

modèle	entraînement	prédiction	AUC	$DETEC_{t_{test}}(\cdot, 500)$
forêt aléatoire	3,08 minutes	2,1 secondes	0,761	205
XGBoost	4,03 minutes	$<$ 0,1 seconde	0,765	219
réseau de neurones	3,65 minutes	$<$ 0,1 seconde	0,764	217

TABLE 2.9 : Performances des modèles

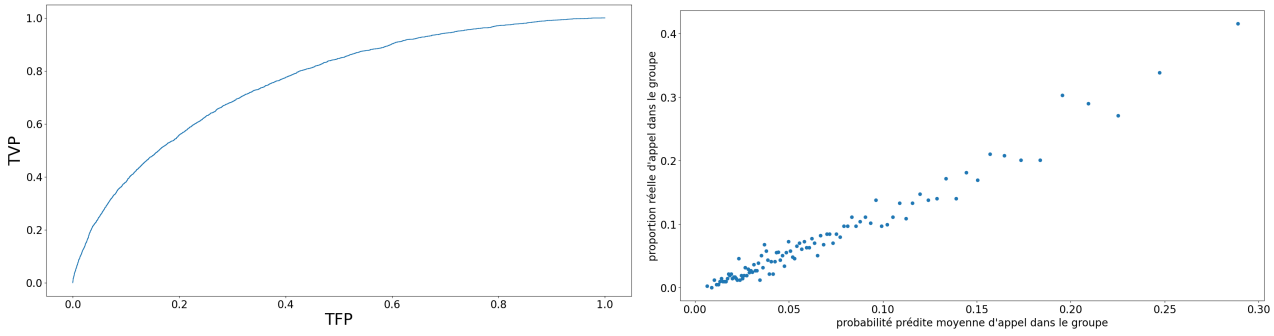


FIGURE 2.20 : Courbe ROC et cohérence du modèle de forêt aléatoire

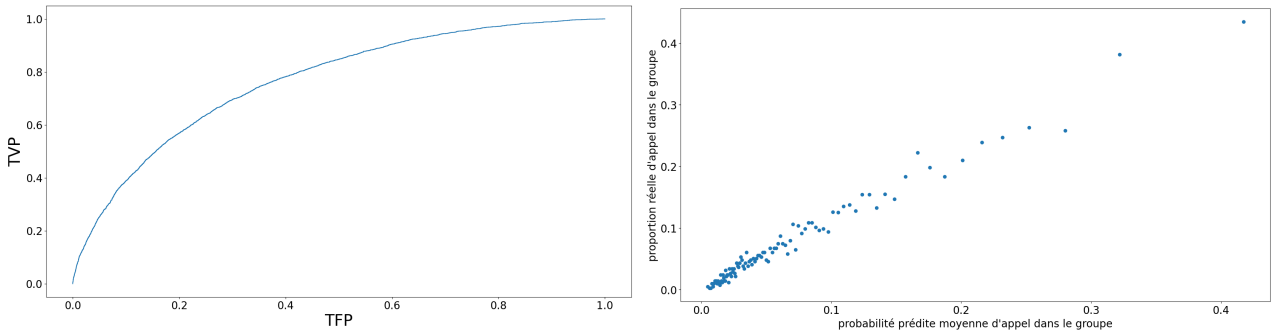


FIGURE 2.21 : Courbe ROC et cohérence du modèle XGBoost

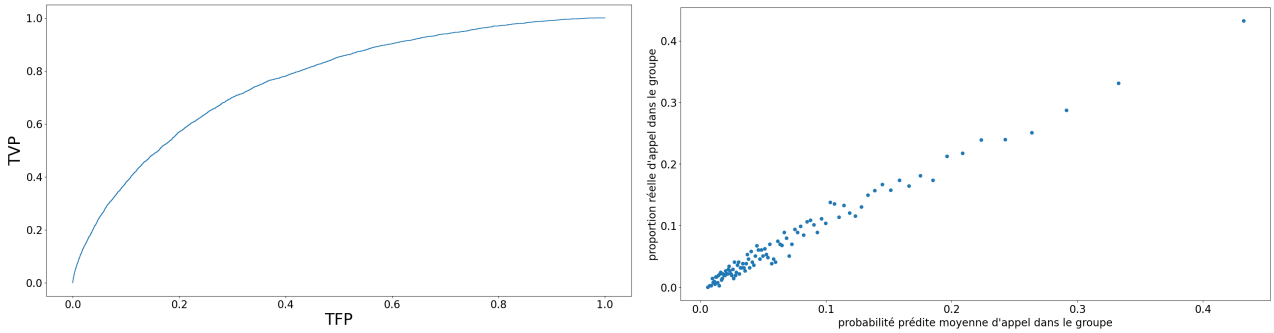


FIGURE 2.22 : Courbe ROC et cohérence du modèle de réseau de neurone

Les performances de ces trois modèles viennent donc améliorer celles des deux modèles plus simples présentés avant tout en conservant des temps de calculs acceptables. Il est possible de constater que les modèles donnent des prédictions cohérentes (notamment le réseau de neurones). Étant donné les natures différentes de ces trois modèles, il est possible d'imaginer qu'ils peuvent se compléter dans la nature des effets qu'ils ont réussi à capter et il peut être intéressant de de les combiner. Si l'on réalise la moyenne des trois estimateurs on obtient une AUC de 0,767 et on parvient à détecter 225 appels à capacité donnée de 500.

2.7 Conclusion

La modélisation de la survenance dans un futur proche d'un appel de gestion est un problème soumis à un fort aléa, d'une part car comme expliqué précédemment nous n'avons pas connaissance de l'ensemble des appels existants dans la réalité mais aussi car le phénomène en lui-même résulte d'un choix de l'assuré. En effet, ce choix peut être orienté par de nombreuses données à disposition de l'assuré mais dont l'assureur n'a pas connaissance comme son emploi du temps, la découverte de nouveaux dégâts liés au sinistre ou même son humeur.

Dans ces conditions, il est raisonnable de considérer les modèles évalués dans ce chapitre (dont les performances sont rappelées dans le tableau 2.10) comme satisfaisants. Il est probable qu'intégrer des données client comme l'âge, le sexe ou encore les précédents dossiers de sinistre conduirait à une amélioration des performances. Cette solution n'a pas été mise en œuvre pour deux raisons principales à savoir une complexité pour accéder à ces informations et le fait que le projet ici est plus de comprendre la dynamique de gestion d'un dossier que les caractéristiques des assurés.

modèle	entraînement	prédiction	AUC	$DETEC_{t_{test}}(\cdot, 500)$
régression logistique	51 secondes	$< 0{,}1$ seconde	0,750	198
arbre de classification	39 secondes	$< 0{,}1$ seconde	0,745	171
forêt aléatoire (fa)	3,08 minutes	2,1 secondes	0,761	205
XGBoost(XGB)	4,03 minutes	$< 0{,}1$ seconde	0,765	219
réseau de neurones (rn)	3,65 minutes	$< 0{,}1$ seconde	0,764	217
$\frac{fa+XGB+rn}{3}$	10,76 minutes	2,2 secondes	0,767	225

TABLE 2.10 : Performances des modèles

Loi agrégée du nombre de dossiers appelants

Une première application des résultats de modélisation peut être la détermination de la loi du nombre de dossiers qui vont générer au moins un appel dans les deux semaines. Si on se place dans une situation de portefeuille clos (on ne considère pas les dossiers ouverts après l'évaluation de la probabilité d'appel) et qu'on note $\mathcal{A}_t$ la loi du nombre de dossiers (parmi ceux ouverts en t) qui vont générer au moins un appel dans les deux semaines on a

$$\mathcal{A}_t = \sum_{d \in D_t} \mathbb{1}_{AG^d_{t+14} > AG^d_t},$$

où les variables dans la somme sont indépendantes. En l'absence de modélisation $\mathcal{A}_t$ suit donc une loi binomiale,

$$\mathcal{A}_t \sim \mathcal{B}(\#D_t, \widehat{p_t}),$$

à supposer qu'on connaisse $\widehat{p_t}$ la probabilité d'appel moyenne. En revanche, si on dispose des probabilités individuelles on a

$$\mathcal{A}_t \sim \sum_{d \in D_t} \mathcal{B}(p^d_t),$$

où les variables dans la somme sont indépendantes. Dans le premier cas on a donc

$$\mathbb{E}(\mathcal{A}_t) = \#D_t \times \widehat{p_t} \text{ et } \mathbb{V}(\mathcal{A}_t) = \#D_t \times \widehat{p_t} \times (1 - \widehat{p_t}),$$

tandis que dans le deuxième on a

$$\mathbb{E}(\mathcal{A}_t) = \sum_{d \in D_t} p^d_t \text{ et } \mathbb{V}(\mathcal{A}_t) = \sum_{d \in D_t} p^d_t \times (1 - p^d_t).$$

Sur notre base de test la modélisation (si on retient la combinaison évoquée en 2.6.2) a permis de réduire l'écart-type de 54,4 à 49,4. Cette légère réduction de l'incertitude sur la charge de gestion peut avoir son intérêt sur l'anticipation des capacités nécessaires. Cependant l'intérêt de la modélisation reste surtout présent à l'échelle du dossier si l'on veut cibler des actions individuelles visant à réduire la charge globale.

Aspect auto-excitant de la gestion

La deuxième conclusion que permet d'apporter ce chapitre est la dimension auto-excitée des dossiers. En effet, on a pu constater que pour la régression logistique et l'arbre de classification δ (le temps écoulé depuis le dernier événement) était le principal déterminant de l'arrivée d'un appel. Ce constat se vérifie aussi sur les trois autres modèles si on calcule l'importance des variables par permutation ou

encore en allant regarder a posteriori la distribution de la prédiction en fonction de δ. Ces résultats viennent appuyer l'observation faite dans la section des statistiques descriptives et il semble donc que le côté "actif" d'un dossier soit primordial pour prédire l'arrivée futur d'un appel.

Cette remarque laisse penser qu'il serait pertinent d'analyser le déroulement du phénomène de gestion dans sa globalité plutôt que de modéliser la survenance d'un événement donné. Le chapitre suivant se concentrera sur cette approche visant à mieux comprendre la succession des événements en gestion de sinistre tout en gardant à l'esprit que les appels de gestions sont les événements dont on cherche le plus à comprendre et prédire la survenance.

Chapitre 3

Modélisation globale de la gestion d'un dossier

La dernière remarque du précédent chapitre conduit à s'intéresser à la dynamique globale de la gestion d'un dossier. Cette approche n'est pas incompatible avec la volonté de mieux comprendre le phénomène d'appel dans la mesure où une compréhension de l'occurrence de l'ensemble des événements ainsi que de leurs interactions devrait être en mesure d'apporter des précisions sur le phénomène particulier d'appel de gestion.

3.1 Idée et objectifs

Dans ce chapitre l'idée est donc d'expliquer et de modéliser la survenance des différents événements d'un dossier au cours du temps. L'ensemble des dossiers ont une trajectoire qui les amène de l'ouverture à la clôture et c'est cette trajectoire, que l'on caractérisera par la survenance d'événements, que l'on va chercher à expliquer. On pose le cadre suivant pour le reste du chapitre.

Soit d un dossier, on fixe sa date d'ouverture comme référence temporelle et on note E l'ensemble des événements possibles. On pose comme dans la définition 2.7 $\left(e_i^d, t_i^d\right)_{i\in\{1,\dots k_d\}}$ l'ensemble des couples $(\textit{événement}, \textit{date})$ du dossier d classé du plus ancien au plus récent avec k_d le nombre total d'événements du dossier d. Si on note T_d le temps que met le dossier d à être clôturé on a donc

$$\left(e_1^d, t_1^d\right) = (\textit{ouverture}, 0) \text{ et } \left(e_{k_d}^d, t_{k_d}^d\right) = (\textit{clôture}, T_d)\,.$$

Soit $e \in E$, on définit à présent le processus de comptage suivant

$$N_t^{e,d} = \sum_{i=1}^{k_d} \mathbb{1}_{\left\{e_i^d = e,\ t_i^d \leq t\right\}} \ \forall t \geq 0. \tag{3.1}$$

On dispose donc d'un processus de comptage par événement et on va donc chercher à modéliser la dynamique suivie par ces processus en sachant que le processus d'ouverture est trivialement constant égal à 1 et que passé le premier saut du processus de clôture tous les processus sont constants. Dans toute la suite du chapitre on travaillera sur l'ensemble des sinistres rentrant dans le cadre de l'étude soit environ 400 000 dossiers (trajectoires) avec en moyenne 8,63 événements (changements d'états) par dossier en comptant l'ouverture et la clôture.

Remarque 3.1. Dans ce chapitre le traitement ou non des appels sera pris en compte car on se focalise davantage sur la compréhension du processus de gestion que sur l'identification d'un besoin de l'assuré.

De manière rigoureuse le processus d'ouverture n'est pas un processus de comptage car ne débutant pas à 0 mais ce détail importe peu car ce processus est déterministe. Les appels de déclaration interviennent uniquement le jour de l'ouverture et ne seront donc pas étudiés sauf mention contraire. La figure 3.1 donne un exemple de trajectoire des processus de chaque événement pour un dossier de notre étude ayant eu une gestion courte et relativement simple alors que la figure 3.2 représente un dossier bien plus lourd en charge de gestion. Le processus de clôture n'a pas été représenté sur ces deux figures pour faciliter leur lecture, la date de clôture correspond à la dernière abscisse affichée.

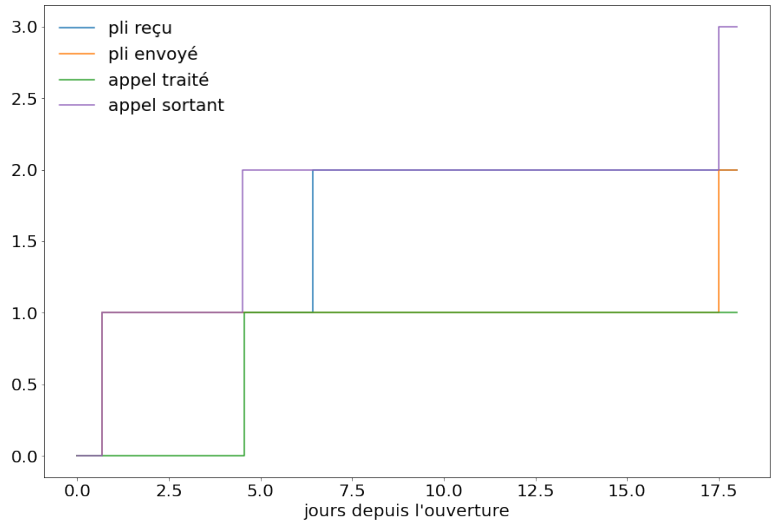


FIGURE 3.1 : Trajectoire de gestion d'un dossier simple

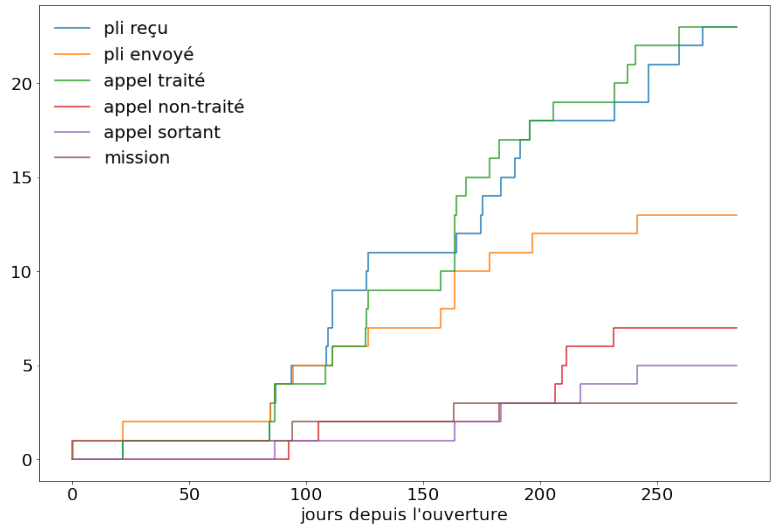


FIGURE 3.2 : Trajectoire de gestion d'un dossier complexe

Ces deux exemples assez opposés laissent entrevoir des périodes d'accalmie communes à tous les processus ainsi que des concentrations de sauts sur des laps de temps très courts. Il est tout à fait logique d'observer ce deuxième phénomène car des plis peuvent être échangés au cours d'un appel ou encore car si un appel est interrompu ou non-traité il est probable qu'un autre le suive peu de temps après.

Une autre façon de représenter la trajectoire des dossiers est l'état en cours du dossier que l'on définit grâce aux temps de sauts $\left(T_n^{e,d}\right)_{n\geq 0}$ que l'on définit par

$$T_n^{e,d} := \inf\{t \geq 0 : N_t^{e,d} \geq n\}.$$

Les temps de sauts correspondent donc bien aux dates des événements introduites ci-dessus et on définit le processus de l'état en cours du dossier E_t^d par

$$E_t^d := \underset{e}{\operatorname{argmax}}\, M_t^d(e), \tag{3.2}$$

avec $M_t^d(e) := \max\left\{T_n^{e,d}, n \geq 0 \mid T_n^{e,d} \leq t\right\}$ le dernier temps de saut de l'événement e intervenu avant t. On peut remarquer que $M_t^d(e) = T^{e,d}_{N_t^{e,d}}$.

Remarque 3.2. Il est important de remarquer que le processus de l'état en cours du dossier ne caractérise pas complètement les processus de comptage associés aux événements. En effet, si deux événements identiques se succèdent le processus de l'état en cours du dossier restera constant après la date du deuxième événement de la même façon que s'il n'avait pas eu lieu.

Le processus $\left(E_t^d\right)_{t\geq 0}$ vit donc dans l'espace E des événements en indiquant le dernier événement connu. Dans la suite modéliser la trajectoire de gestion d'un dossier se fera donc en modélisant $\left(E_t^d\right)_{t\geq 0}$ ou en modélisant simultanément les $\left(N_t^{e,d}\right)_{t\geq 0}$, $e \in E$ en gardant à l'esprit la remarque 3.2.

3.2 Approche par processus de Markov

Une première approche qui va être évaluée est celle d'une modélisation par processus de Markov. On va donc chercher à modéliser les transitions du processus $\left(E_t^d\right)_{t\geq 0}$ que ce soit en termes probabilité de transition d'un événement à un autre ou en termes de temps d'attente. Le processus de Markov est une modélisation simple d'un processus à saut à temps continu est servira donc de première approche.

3.2.1 Cadre théorique

Les éléments de cette section sont adaptés de Norris (1997). Pour simplifier les notations on assimile E à $\{1, \ldots, n\}$ (avec $n = \#E$) l'ensemble fini des états possibles du processus $\left(E_t^d\right)_{t\geq 0}$. On suppose que ce dernier est un processus markovien de sauts.

Définition 3.1. On dit que $\left(E_t^d\right)_{t\geq 0}$ est un processus markovien de sauts si

$$\mathbb{P}\left(E_{s+t}^d = j \,\middle|\, E_s^d = i, E_{s_{n-1}}^d = i_{n-1}, E_{s_{n-2}}^d = i_{n-2}, \ldots, E_{s_0}^d = i_0\right) = \mathbb{P}\left(E_{s+t}^d = j \,\middle|\, E_s^d = i\right),$$

pour tout $0 \leq s_0 < s_1 < \cdots < s_{n-2} < s_{n-1} < s < s+t$ et tout $i_0, i_1, \ldots, i_{n-2}, i_{n-1}, i, j \in E$ tels que le membre de gauche soit bien défini.

On dit de plus que le processus est homogène en temps si $\mathbb{P}\left(E_{s+t}^d = j \,\middle|\, E_s^d = i\right)$ ne dépend pas de s et on note $\mathbb{P}\left(E_{s+t}^d = j \,\middle|\, E_s^d = i\right) = P_t(i,j)$. Dans la suite on suppose également que $\left(E_t^d\right)_{t\geq 0}$ est homogène en temps et on a la propriété suivante.

Proposition 3.1 (Équation de Chapman-Kolmogorov)**.** Les matrices $(P_t)_{t\geq 0}$ vérifient la relation suivante

$$P_{s+t} = P_s P_t.$$

Cette propriété montre que P_t est connue pour tous les temps t si on la connaît pour t petit et peut donc être déterminée par sa dérivée (à droite) et $t = 0$ au sens de la proposition suivante.

Proposition 3.2. Il existe une matrice $(Q_{ij})_{i,j\in E}$ (appelée le générateur infinitésimal) qui vérifie

$$Q_{ij} \geq 0 \text{ si } i \neq j \text{ et } Q_{ii} = -\sum_{j\neq i} Q_{ij},$$

et telle que, lorsque $t \searrow 0$

$$P_h(i,j) = hQ_{ij} + o(h) \text{ si } i \neq j \text{ et } P_h(i,i) = 1 + hQ_{ii} + o(h).$$

On pose T_n comme le n-ième temps de saut de $\left(E_t^d\right)_{t\geq 0}$

$$T_n := \inf\left\{t > T_{n-1} \mid E_t^d \neq E_{T_{n-1}}^d\right\},\ T_0 := 0.$$

Si on ne regarde maintenant plus que les états parcourus par le processus $\left(E_t^d\right)_{t\geq 0}$, on a la propriété suivante.

Proposition 3.3. En posant le processus discret $Z_n = E_{T_n}^d$ avec $(T_n)_{n\geq 0}$ les temps de saut de $\left(E_t^d\right)_{t\geq 0}$ on a que $(Z_n)_{n\geq 0}$ est une chaîne de Markov (en temps discret) de matrice de transition Π telle que

$$\Pi_{ij} = \begin{cases} -\frac{Q_{ij}}{Q_{ii}} & \text{si } i \neq j \text{ et } Q_{ii} \neq 0, \\ 0 & \text{si } i \neq j \text{ et } Q_{ii} = 0, \end{cases}$$
$$\Pi_{ii} = \begin{cases} 0 & \text{si } Q_{ii} \neq 0, \\ 1 & \text{si } Q_{ii} = 0. \end{cases}$$

On peut montrer que les temps d'inter-arrivées $(\delta_n)_{n\geq 1} := (T_n - T_{n-1})_{n\geq 1}$ sont indépendants et suivent des lois exponentielles conditionnellement à $(Z_n)_{n\geq 0}$ dont les intensités sont données par $-Q_{Z_{n-1}Zn}$. On peut donc caractériser le processus markovien de sauts homogène en temps $\left(E_t^d\right)_{t\geq 0}$ grâce à sa matrice de générateur infinitésimal $(Q_{ij})_{i,j\in E}$ par le schéma suivant.

- Sachant que le processus démarre en $E_0^d = i_0$ il y reste pendant un temps δ_1 exponentiel de paramètre $-Q_{i_0i_0}$ puis saute dans un nouvel état i_1 avec probabilité de $-\frac{Q_{i_0i_1}}{Q_{i_0i_0}}$ pour $i \neq j$.
- ...
- Arrivé en i_n le processus y reste un temps δ_n exponentiel (indépendant des autres) de paramètre $-Q_{i_ni_n}$ puis saute dans un nouvel état i_{n+1} avec probabilité de $-\frac{Q_{i_ni_{n+1}}}{Q_{i_ni_n}}$ pour $i_{n+1} \neq i_n$.

Ce schéma caractérise bien le processus $\left(E_t^d\right)_{t\geq 0}$ si l'on considère qu'une loi exponentielle d'intensité nulle est infinie et que donc le processus devient constant s'il rencontre ce cas. Étant établi ce schéma il est possible de prendre en compte la remarque 3.2 et de légèrement adapter la modélisation pour qu'elle prenne en compte les successions d'événements identiques. Pour cela il suffit de laisser la possibilité au processus de sauter sur place avec une certaine probabilité et on pose un modèle légèrement différent en supposant que les événements suivent la dynamique suivante.

- À n'importe quel instant la nature du prochain événement ne dépend que de celle de l'événement précédent selon une matrice de transition $\tilde{\Pi}$.

- À n'importe quel instant le temps d'attente jusqu'au prochain événement ne dépend que de la nature de l'événement précédent et suit une loi exponentielle d'une intensité déterminée par cette nature et on note $\Lambda = (\lambda_1, \ldots, \lambda_n)$ ces intensités.

La différence avec un processus markovien de sauts homogène en temps vient du fait qu'il est possible d'avoir simultanément $\tilde{\Pi}_{ii} > 0$ et $\lambda_i > 0$. Si on considère les événements d'ouverture et de clôture comme respectivement associés aux valeurs 1 et n, alors on a

$$\begin{aligned} \tilde{\Pi}_{i1} &= 0 \quad \forall i \in E \\ \tilde{\Pi}_{ni} &= 0 \quad \forall i \neq n \\ \tilde{\Pi}_{nn} &= 1 \\ \lambda_i &> 0 \quad \forall i \neq n \\ \lambda_n &= 0 \end{aligned}$$

et on considère aussi logiquement que le processus démarre par l'ouverture donc par l'événement 1 en $t = 0$.

Calibration des paramètres

En considérant que l'ensemble des dossiers $d \in D$ sont des réalisations indépendantes de la dynamique décrite ci-dessus alors l'ensemble des transitions observées sont indépendantes et un estimateur de $\tilde{\Pi}$ est donc donné par

$$\tilde{\Pi}_{ij} \simeq \frac{\text{Nombre d'observations où l'événement j succède à l'événement i}}{\text{Nombre d'observations de l'événement i}}.$$

De même les temps d'inter-arrivées observés sont indépendants et si on pose $\left(d_k^i\right)$ l'ensemble des temps d'inter-arrivées observés après i alors un estimateur de Λ est donc donné par

$$\lambda_i \simeq \frac{\text{Nombre d'observations de l'événement i}}{\sum_k d_k^i}.$$

Ces estimateurs sont à la fois des estimateurs par la méthode des moments mais correspondent également à une estimation par maximum de vraisemblance.

3.2.2 Application et interprétation

Les estimations ont été réalisées sur l'ensemble des sinistres qui composent la base d'étude et de ce fait pour ceux qui ne sont pas encore clôturés on ne prendra pas en compte le dernier événement dans les estimations. Le tableau 3.1 donne les estimations des probabilités de transition où "n-t", "t" et "s" signifient respectivement non-traité, traité et sortant (on rappelle que les appels de déclaration ne sont pas pris en compte). Les deux dernières lignes du tableau donnent les estimations des intensités de saut ainsi que leur inverse pour plus de lisibilité (une loi exponentielle ayant pour espérance l'inverse de son intensité).

Il est possible de lire le tableau de la façon suivante. À l'ouverture le prochain événement arrive après un temps d'attente exponentiel d'intensité 0,223 soit en moyenne au bout de 4,48 jours et cet événement est un pli reçu dans 35,5% des cas par exemple. Certains constats sont assez évidents comme par exemple le fait que l'ouverture d'un ordre de mission est rapidement suivie d'une nouvelle interaction. Il est en effet peu probable que le gestionnaire décide d'acter qu'un expert doive estimer le montant d'un sinistre sur place sans que cela ne donne lieu à un quelconque échange avec l'assuré. En revanche, certains faits interpellent et c'est notamment le cas en regardant la colonne des probabilités d'arrivée sur l'événement d'appel traité. Si on excepte les appels non traités (on ne regarde donc pas

les gens qui rappellent suite à une tentative infructueuse d'appeler l'assureur) la valeur la plus élevée de la colonne est atteinte pour la ligne correspondant également à l'appel traité. Cela s'interprète de la façon suivante, hormis le phénomène de rappel, l'événement qui a le plus de chance de "générer" un appel traité est l'appel traité lui-même ce qui peut sembler assez étrange. En effet, deux appels de gestion traités successifs peuvent laisser penser que la situation du dossier n'a pas particulièrement évolué entre les deux et ces appels sont probablement au moins en partie des appels "inutiles". Cette dernière remarque est d'autant plus marquée quand on prend en compte le fait que seuls les appels entrants peuvent être manquants de manière significative dans nos données.

événement	ouverture	appel n-t	appel t	appel s	mission	pli envoyé	pli reçu	clôture
ouverture	0,000	0,022	0,038	0,107	0,101	0,251	0,355	0,126
appel n-t	0,000	0,223	0,387	0,058	0,023	0,055	0,194	0,060
appel t	0,000	0,063	0,173	0,074	0,019	0,266	0,265	0,141
appel s	0,000	0,064	0,066	0,166	0,016	0,322	0,172	0,194
mission	0,000	0,047	0,140	0,169	0,007	0,498	0,129	0,010
pli envoyé	0,000	0,035	0,053	0,115	0,017	0,135	0,290	0,355
pli reçu	0,000	0,034	0,054	0,175	0,078	0,223	0,353	0,083
clôture	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000
intensité	0,223	0,228	0,101	0,114	0,685	0,063	0,117	0,000
1/intensité	4,48	4,39	9,88	8,75	1,46	15,77	8,58	-

TABLE 3.1 : Valeurs estimées de la matrice de transition et des intensités

Pertinence du modèle

Si l'on cherche à évaluer la pertinence du modèle il est légitime d'essayer de faire cette validation par des tests statistiques. Il existe des tests permettant de tester si un processus est bien markovien (notamment en étudiant la chaîne de Markov telle que définie à la proposition 3.3) mais ces tests reposent sur une observation sur le temps long du processus. Dans le cas étudié on dispose d'un grand nombre d'observation du processus mais qui s'arrête au moment de la clôture. Il est tout de même possible de procéder à des tests statistiques sur le modèle posé en testant par exemple le caractère exponentiel des temps d'inter-arrivées. En procédant de la sorte on suppose que tous les temps d'inter-arrivée en provenance d'un événement sont indépendants et identiquement distribués et on test l'hypothèse que leur loi est bien une loi exponentielle. Le test de Kolmogorov-Smirnov permet de tester l'adéquation a une loi en évaluant la distance de la fonction de répartition empirique d'un échantillon à la fonction de répartition de cette loi candidate. En se fixant un seuil à 5% le test rejette l'hypothèse d'adéquation pour les temps d'inter-arrivées en provenance de chacun des événements.

Au-delà de ce test le modèle semble être trop restrictif pour modéliser les phénomènes de périodes d'excitation ou de calme évoqués en début de chapitre qui semblent peu compatibles avec une loi sans mémoire. Si l'on considère la modélisation de la probabilité d'appel du chapitre précédent en considérant comme valide le modèle de cette section on aurait obtenu que la probabilité d'appel d'un dossier au moment du scan est complètement déterminée par la nature du dernier événement connu (et pas par le temps écoulé depuis car la loi exponentielle est sans mémoire). Dans la pratique un tel modèle pour la probabilité d'appel donne une AUC de 0,607 donc bien inférieure à celles des modèles évalués au chapitre précédent ce qui vient appuyer le rejet des hypothèses faites sur la dynamique du processus de gestion dans cette section bien que ce dernier ne soit pas complètement inutile dans la mesure où les informations apportées dans le tableau 3.1 permettent de tirer des enseignements.

3.3 Approche par processus de Hawkes

Dans cette section on s'intéressera à une classe particulière de processus de comptage qui permettent de tenir compte des événements passés de façon que l'arrivée d'un événement viennent temporairement augmenter la probabilité d'un autre événement. On fait ici référence aux processus de Hawkes, HAWKES (1971), qui trouvent de nombreuses applications pour la modélisation de phénomènes auto-excités dans des domaines variés comme l'épidémiologie, la finance, l'étude des répliques sismiques ou plus récemment en assurance pour modéliser le risque cyber ou terroriste.

3.3.1 Définitions

Les informations de cette section sont adaptées de Bacry et al. (2015). Les processus de Hawkes permettent de capter l'auto-excitation de processus de comptage au travers de la notion d'intensité d'un processus de comptage. On se place dans un premier temps dans le cadre d'un processus de comptage univarié avant d'introduire des processus en dimensions supérieures.

Définition 3.2 (Intensité conditionnelle)**.** Soit $(N_t)_{t>0}$ un processus de comptage tel que

$$N_t = \sum_{n\geq 0} \mathbb{1}_{T_n<t},$$

avec (T_n) les temps de saut du processus. On définit le processus d'intensité conditionnelle associé à $(N_t)_{t>0}$ noté λ par

$$\lambda(t) = \lim_{h\to 0} \mathbb{E}\left(\frac{N_{t+h} - N_t}{h} \middle| \mathcal{F}_t \right),$$

où la filtration $\mathcal{F}_t$ représente l'information connue sur le processus jusqu'au temps t exclu.

L'intensité conditionnelle d'un processus représente donc la force avec laquelle le processus saute en t au sens où pour $\mathrm{d}t$ petit le processus sautera sur $[t, t+\mathrm{d}t]$ avec probabilité $\lambda(t)\mathrm{d}t$. Le processus de Hawkes univarié est un processus de comptage dont l'intensité dépend des temps de sauts passés au sens suivant.

Définition 3.3 (Processus de Hawkes univarié)**.** Un processus de Hawkes univarié est un processus de comptage dont l'intensité conditionnelle adopte la forme suivante

$$\lambda(t) = \mu + \int_0^t \phi(t-s)\mathrm{d}N_s = \mu + \sum_{T_n<t} \phi(t-T_n),$$

avec $\mu \geq 0$ l'intensité de base et $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ la fonction noyau qui caractérise l'auto-excitation.

L'auto-excitation vient du fait que la survenance de sauts vient augmenter le nombre de terme dans la somme. On considère parfois des processus de Hawkes où l'intensité de base n'est pas constante mais une fonction du temps ce qui peut être utile pour modéliser une évolution du comportement général comme une périodicité. Le noyau permet de déterminer comment évolue l'impact d'un saut sur la survenance d'autres sauts au cours du temps, s'il est décroissant un saut augmente de moins en moins l'intensité au fur et à mesure que le temps s'écoule. On peut remarquer que le choix du noyau nul correspond simplement à un processus de Poisson. L'exemple de noyau le plus souvent utilisé de par sa simplicité et des avantages en termes de calibration des paramètres est le noyau exponentiel qui s'écrit sous la forme

$$\phi(s) = \alpha\beta e^{-\beta s},$$

avec $\alpha, \beta > 0$. La figure 3.3 donne un exemple de simulation d'un processus de Hawkes à noyau exponentiel ($\alpha = 0,5, \beta = 0,1$) avec une intensité de base $\mu = 0,2$. La courbe représente l'intensité et les points les temps de sauts.

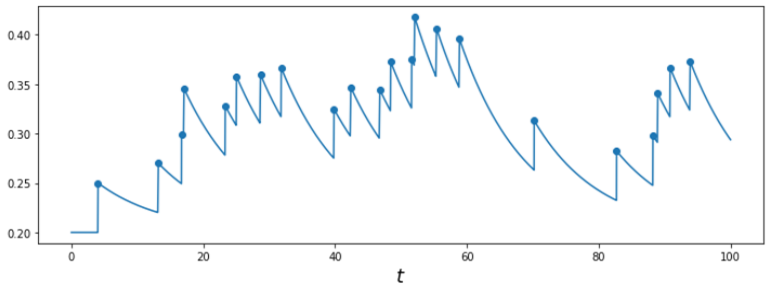


FIGURE 3.3 : Intensité et temps de sauts d'un processus de Hawkes à noyau exponentiel

Dans le cas de la modélisation du processus de gestion l'intérêt d'une approche par processus de Hawkes serait donc de prendre en compte le fait que la survenance d'un événement augmente momentanément la probabilité d'une nouvelle survenance d'événement. Étant donné que le processus de gestion se compose de plusieurs processus de comptage on introduit les processus de Hawkes multivariés qu'on peut définir en généralisant la définition 3.3 de la manière suivante.

Définition 3.4 (Processus de Hawkes multivarié)**.** Soient $\left(N_t^1\right)_{t>0}, \ldots, \left(N_t^d\right)_{t>0}$ d processus de comptage (on note t_n^i les temps de sauts respectifs), on dit qu'ils forment un processus de Hawkes multivarié si les intensités conditionnelles adoptent la forme suivante

$$\lambda^i(t) = \mu^i + \sum_{j=1}^{d} \int_0^t \phi_{ij}(t-s)\mathrm{d}N_s^j = \mu^i + \sum_{j=1}^{d} \sum_{T_n^j < t} \phi_{ij}(t - T_n^j),$$

avec $\mu^i \geq 0$ les intensités de base et $\phi_{ij} : \mathbb{R}^+ \to \mathbb{R}^+$ les fonctions noyaux et on note Φ la matrice noyau qui en est composée.

Le principe reste le même qu'en dimension 1 mais la survenance d'un type d'événement vient impacter la survenance future des événements des autres types d'événements en plus de lui-même. Comme dans le cas univarié on considère parfois le cas d'intensités de base non-constante. On définit le noyau exponentiel dans le cas multivarié par

$$\phi_{ij}(s) = \alpha_{ij}\beta_{ij}e^{-\beta_{ij}s},$$

avec $\alpha_{ij}, \beta_{ij} > 0, 1 \leq i, j \leq d$. Les paramètres peuvent s'interpréter de la manière suivante, α_{ij} indique avec quel force un saut du processus j vient perturber le processus i tandis que β_{ij} indique la manière dont cette perturbation va s'étaler dans le temps.

La figure 3.4 donne un exemple de simulation d'un processus de Hawkes en deux dimensions avec un noyau exponentiel tel que $\alpha_{11} = \alpha_{22} = 0.2$, $\alpha_{12} = \alpha_{21} = 0.4$, $\beta_{11} = \beta_{22} = \beta_{12} = \beta_{21} = 0.3$ et $\mu_1 = \mu_2 = 0.1$. On peut y observer les interactions entre les différents temps de sauts par le biais des intensités. On retrouve bien des périodes de calmes et de forte activité ce qui est une propriété souhaitée dans la modélisation du processus de gestion. L'emballement constaté entre les temps $t = 60$ et $t = 70$ laisse imaginer qu'un processus de Hawkes puisse rentrer dans un état de dérive où l'intensité s'accumule entraînant une arrivée fréquente de saut entraînant elle-même une augmentation de l'intensité et ainsi de suite de tel sorte que le processus explose. Les conditions pour ne pas être dans ce cas pathologique sont données par la proposition 3.4.

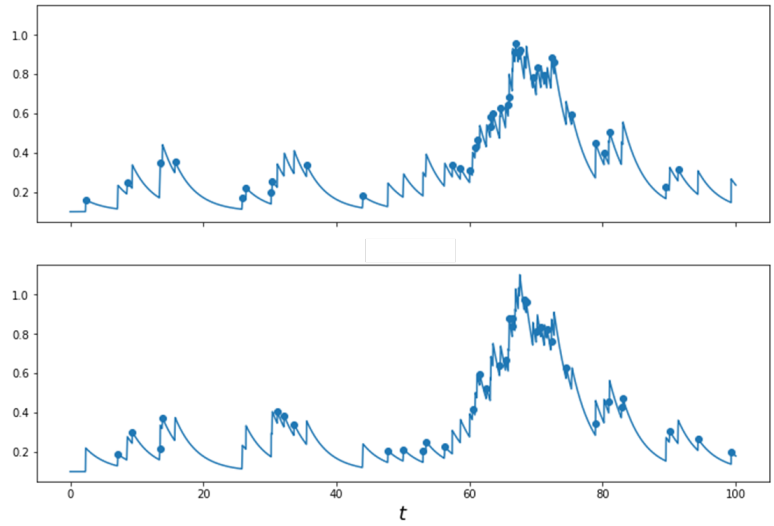


FIGURE 3.4 : Intensité et temps de sauts d'un processus de Hawkes à noyau exponentiel

Proposition 3.4 (Stabilité)**.** On dit qu'un processus de Hawkes est stable s'il est asymptotiquement à accroissements stationnaires et que son intensité est asymptotiquement stationnaire. Un processus de Hawkes est stable si son noyau vérifie la condition suivante

$$\|\Phi\| := \left\{ \int_0^{+\infty} \phi_{ij}(s)\mathrm{d}s \right\}_{i,j=1,\dots,d} \text{ est une matrice ayant un rayon spectral } < 1.$$

Dans le cas d'un noyau exponentiel cette condition revient donc à ce que la matrice composée des α_{ij} ne possède pas de valeur propre de module supérieur à 1.

La calibration d'un processus de Hawkes multivarié peut être une tâche compliquée à réaliser mais peut se faire par principe du maximum de vraisemblance. Dans le cas d'un noyau exponentiel l'expression de la vraisemblance du modèle permet de procéder à son optimisation de manière plus rapide que pour d'autre noyau ce qui est une des raisons de la popularité de ce choix de noyau. Pour un processus de Hawkes de dimension d la log-vraisemblance est donnée par

$$\log \mathcal{L}(\mu, \Phi) = -\sum_{i=1}^{d} \int_0^T \lambda^i(t)\mathrm{d}t + \sum_{i=1}^{d} \sum_{m=1}^{M_i} \log \lambda^i(t_m^i),$$

où T est le temps sur lequel est observé le processus et $\left(t_m^i\right)_{m=1,\dots,M^i}$ sont les temps de sauts de la i-ème composante du processus.

Une propriété intéressante des processus de Hawkes est qu'ils peuvent être vus avec une structure de processus de branchement en considérant le procédé de simulation suivant.

- Pour $i = 1, \dots, d$ on tire des événements de génération 0 $\left(t_m^{i,0}\right)_{m=1,\dots,M^{i,0}}$ selon un processus de poisson homogène en temps d'intensité μ^i sur l'intervalle $[0, T]$.
- Pour tout événement de type j arrivé en $t_{m'}^{j,0}$ et pour $i = 1, \dots, d$ on tire des événements de génération 1 $\left(t_m^{i,1}\right)_{m=1,\dots,M^{i,1}}$ selon un processus de poisson non-homogène en temps d'intensité $\phi_{ij}(t - t_m'^{j,0})$ sur l'intervalle $[t_m'^{j,0}, T]$.

- On génère une nouvelle génération jusqu'à ce que plus aucun événement ne soit généré sur $[0, T]$.

La réunion de tous les événements tirés dans la procédure ci-dessus correspond à la réalisation d'un processus de Hawkes multivarié tel que défini en 3.4. Cette représentation est particulièrement intéressant car la structure de descendance donne un lien de causalité entre les événements survenus. En pratique on a donc qu'en moyenne un événement de type j donne naissance à ν_{ij} événement de type i avec $\nu_{ij} = \int_0^{+\infty} \phi_{ij}(s)\mathrm{d}s$. Dans le cas particulier d'un noyau exponentiel $\nu_{ij} = \alpha_{ij}$ et la "natalité" des événements est directement un paramètre du modèle.

3.3.2 Application et interprétation

On suppose maintenant que les processus de comptage associés aux événements (hors clôture et ouverture) de la gestion d'un dossier sont la réalisation d'un processus de Hawkes à noyau exponentiel dont on observe la réalisation jusqu'à la clôture. L'estimation des paramètres a été réalisée grâce à la librairie Python *tick*, Bacry et al. (2017), face a des problèmes de temps de calibration l'hypothèse de coefficients $(\beta_{ij})_{i,j=1,\dots,d}$ constants a été retenue. On obtient après calibration un coefficient $\beta = 1,39$ qui peut s'interpréter de la sorte, l'influence d'un événement sur la survenance d'autres événements n'a plus que $e^{-1,39} \simeq 0,24 = 24\%$ de son impact initial au bout d'un jour. On donne la transposée de la matrice des coefficients α_{ij} et des intensités de base μ^i dans le tableau 3.2.

événement	appel n-t	appel t	appel s	mission	pli envoyé	pli reçu	intensité de base
appel n-t	0,2915	0,3459	0,2824	0,2817	0,3137	0,3222	0,0012
appel t	0,0425	0,0645	0,0413	0,0418	0,0314	0,1015	0,0039
appel s	0,0290	0,0449	0,0236	0,0151	0,0162	0,0638	0,0063
mission	0,0484	0,0655	0,0771	0,0323	0,7538	0,1060	0,0031
pli envoyé	0,0000	0,0005	0,0000	0,0000	0,0000	0,0728	0,0187
pli reçu	0,0000	0,0011	0,0216	0,0181	0,0708	0,0269	0,0257

Table 3.2 : Valeur estimées de la transposée de la matrice des α_{ij} et des intensités de base μ^i

La valeur du paramètre β impose d'être prudent sur les interprétations données au modèle. En effet, une décroissance si importante de l'intensité provoquée par un saut suggère que le processus n'a capté des interactions que de court terme (à l'échelle du jour). Une fois cette remarque prise en compte les résultats peuvent tout de même mettre en avant certains phénomènes comme par exemple le fait qu'une ouverture d'ordre de mission engendre 0,75 plis envoyés ce qui peut sembler assez pertinent dans la mesure où l'envoi de plis à ce moment est une pratique standardisée. Au niveau des appels traités on peut constater que sur des interactions de court terme ce sont largement les appels non traités qui en génèrent le plus (0,35) et ce phénomène correspond donc probablement simplement aux assurés qui rappellent instantanément ou presque après une tentative infructueuse.

Une des raisons qui ont pu conduire à une modélisation ne prenant en compte qu'une auto-excitation à court terme est le fait que l'on dispose de beaucoup de trajectoires mais que celles-ci ne comportent que quelques sauts. La présence d'événement rapprochés vient donc donner un poids très important à l'auto-excitation de court terme car elle ne peut pas être compensée par de nombreuses observations espacées dans le temps. Le fait de considérer qu'après un laps de temps court les événements ne sont plus générés que par l'intensité de base est une limite dans notre modélisation car on se rapproche du cas markovien étudié à la section précédente.

On pourrait essayer de capter des interactions de plus long terme en relâchant l'hypothèse de constance des β_{ij} ou en choisissant des noyaux plus adaptés à ce genre de tâches mais on se heurte à

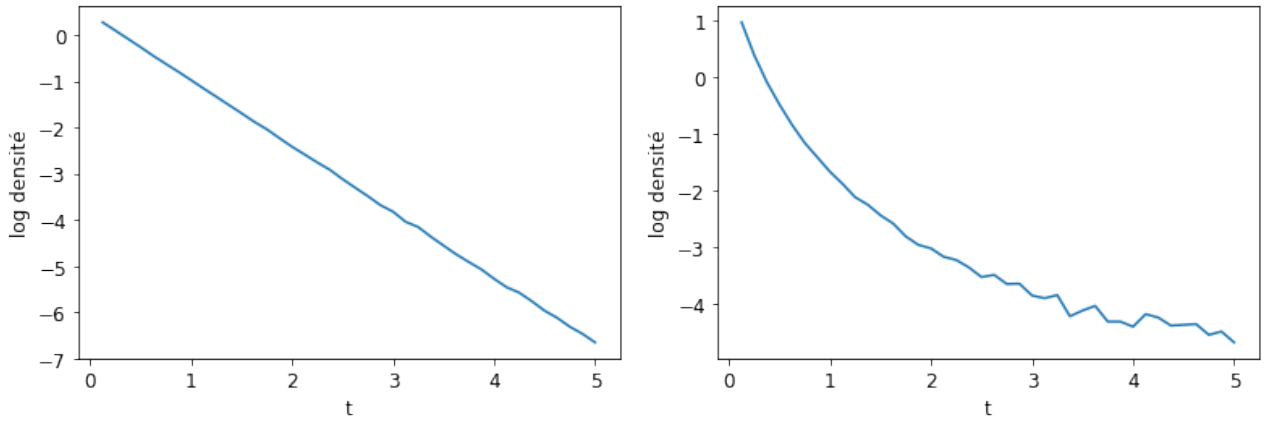
la problématique de calibrer un processus de Hawkes sur beaucoup de "petites" trajectoires. Dans la pratique les processus de Hawkes sont souvent calibrés sur des trajectoires longues pour être pertinents.

3.4 Utilisation de l'intensité comme score

L'impossibilité de parvenir à calibrer un processus de Hawkes de manière satisfaisante dans l'optique de modéliser une auto-excitation n'enlève pas la présence de ce comportement. Il va être ici développée une approche s'inspirant de l'intensité d'un processus de Hawkes à noyau exponentiel pour l'appliquer à la modélisation de probabilité d'appel (problème du chapitre précédent).

3.4.1 Distribution des temps d'inter-arrivée

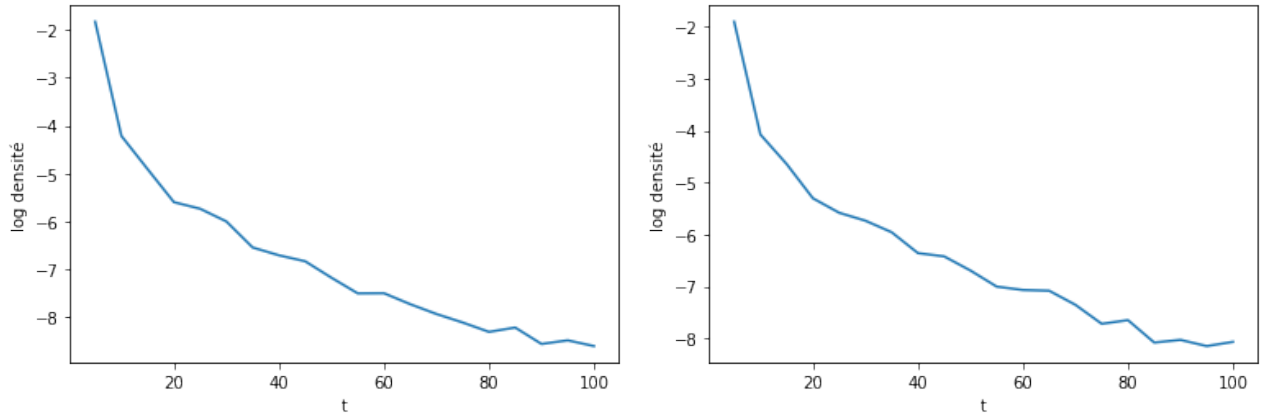
Une façon de se convaincre de la présence d'un phénomène d'auto-excitation est d'étudier les temps d'inter-arrivées. Dans le cas d'un processus markovien ces temps sont distribués de façon exponentielle et en présence d'une auto-excitation on s'attend donc à avoir des distributions plus concentrées en 0 soit donc à décroissance plus forte que pour une loi exponentielle. En effet, on souhaite que l'intensité avec laquelle le processus saute décroisse entre les sauts comme c'est le cas pour pour les processus de Hawkes. On peut facilement observer la différence entre les deux situations en calculant le logarithme de la densité des temps d'inter-arrivées qui doit être linéaire dans le cas d'un processus markovien et convexe dans le cas où les sauts seraient "groupés" du fait d'une auto-excitation. La figure 3.5 illustre bien ce constat en donnant le logarithme de la densité des temps d'inter-arrivées pour un processus de Poisson et un processus de Hawkes univarié à noyau exponentiel.



(a) Processus de Poisson (b) Processus de Hawkes à noyau exponentiel

FIGURE 3.5 : Logarithme de la densité des temps d'inter-arrivées sur simulations

Si l'on revient maintenant sur les processus de comptage des événements de la gestion d'un dossier on aimerait observer des temps d'inter-arrivées suivant une distribution faisant apparaître une auto-excitation. La figure 3.6 donne deux exemples de logarithmes de densités de temps d'inter-arrivées entre deux types d'événements. Les appels traités et non-traités ont à nouveau été rassemblés car on va par la suite chercher à se rapprocher du cadre du chapitre 2 et on représente les temps d'inter-arrivées qui aboutissent à un appel entrant en provenance d'un appel sortant ou entrant. Il est possible sur ces deux exemples d'observer une convexité qui laisse à penser que l'on ne se trouve pas dans une situation de lois sans mémoire mais plutôt dans des cas de distribution pouvant correspondre à un phénomène d'auto-excitation.



(a) ... en partant d'un appel entrant

(b) ... en partant d'un appel sortant

FIGURE 3.6 : Logarithme de la densité des temps d'inter-arrivées aboutissants à un appel entrant...

3.4.2 Retour sur la modélisation du phénomène d'appel

Le constat précédent met en avant que si la modélisation par un processus de Hawkes s'est montrée globalement infructueuse, les événements du processus de gestion présentent des similarités avec la répartition des temps de sauts d'un processus de Hawkes. Si on reprend la notion d'intensité conditionnelle cela suggère que cette dernière ne soit pas constante. En cherchant à adapter la notion d'intensité conditionnelle au problème de modélisation de la probabilité d'appel du chapitre précédent on peut faire l'approximation suivante. Si le processus de gestion avait une certaine intensité d'appel entrant $\lambda^*(t)$ à l'instant t, alors si l'on considère l'intervalle de 14 jours comme petit la probabilité d'appel devrait être proportionnelle à $\lambda^*(t)$. Selon ce principe et en s'inspirant de la forme paramétrique du noyau exponentiel on pose le modèle suivant,

$$a_t^d = \mathbb{P}(AG_{t+14}^d > AG_t^d \mid \mathcal{F}_t^d) = f\left(\mu + \sum_{j\in E}\sum_{T_n^j<t} \alpha_j\beta_j e^{-\beta_j(t-T_n^j)}\right), \tag{3.3}$$

avec E l'ensemble des événements, $\left(T_n^j\right)_{n\geq 0}$ les temps d'arrivées des événements de type j et f une fonction croissante à valeurs dans $[0,1]$. On peut considérer μ comme une intensité de base d'appel de gestion, α_j comme la force globale avec laquelle un événement de type j vient provoquer des appels de gestion et β_j comme caractérisant la décroissance de cet effet. L'utilité de la fonction f est de ramener l'intensité dans l'intervalle $[0,1]$ pour l'interpréter comme une probabilité. Dans la mesure où l'on choisit une fonction croissante, une fois les paramètres fixés le choix de f n'impactera pas nos métriques de performance dans la mesure ou l'AUC et $DETEC_t(\cdot, 500)$ ne dépendent que de l'ordre dans lequel sont classés les dossiers par le modèle. On peut choisir la fonction sigmoïde inverse du lien *logit*, soit $\sigma(x) := \frac{1}{1+e^{-x}}$, pour représenter l'intensité comme une probabilité.

Il est intéressant de constater que lorsqu'on fige les valeurs des β_j on se retrouve en présence d'un modèle de régression logistique. En effet, si on pose $X_j = \sum_{T_n^j<t} \beta_j e^{-\beta_j(t-T_n^j)}$ on peut alors réécrire l'équation (3.3) comme

$$a_t = \sigma\left(\mu + \sum_{j\in E} \alpha_j X_j\right), \tag{3.4}$$

qui correspond bien à un modèle de régression logistique. Il est donc possible de voir la modélisation proposée comme une régression logistique possédant des hyper-paramètres $(\beta_j)_{\in E}$ qui viennent définir les variables prédictives utilisées. Ce constat va être utile pour calibrer les paramètres du modèle, on cherche les variables prédictives et donc les $(\beta_j)_{\in E}$ qui maximisent les performances de la régression logistique. Si dans un premier temps on considère que tous les effets décroissent à la même vitesse on obtient $\beta_j = \beta \ \forall j \in E$ et on cherche juste à maximiser les performances de la régression logistique en fonction en fonction du paramètre β. Si l'on prend l'AUC pour métrique et que l'on tient pour acquis la calibration d'une régression logistique il ne reste plus qu'à maximiser la fonction qui associe l'AUC du modèle à β. Une représentation de cette fonction sur l'intervalle $[0, 2]$ est donnée à la figure 3.7 et elle semble clairement être unimodale, il est donc possible de maximiser la fonction par dichotomie. Sur l'ensemble d'entraînement on trouve $\beta^{max} = 0.04$ et donc le modèle semble capter des interactions d'auto-excitation sur un temps bien plus long que lors de la tentative de calibrer un processus de Hawkes. En effet, l'impact d'un saut ne perdra la moitié de son impact qu'après $-\frac{\log(0.5)}{0.04} \simeq 17$ jours ce qui semble plus satisfaisant.

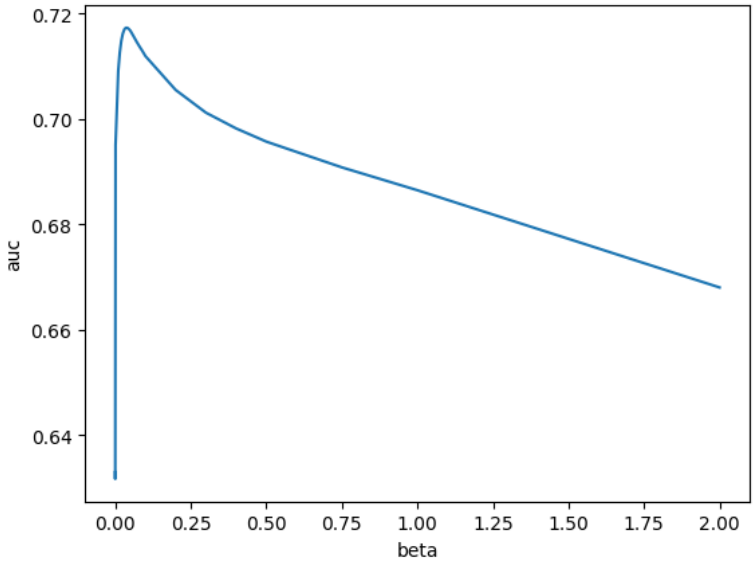


FIGURE 3.7 : AUC du modèle en fonction du choix de β

Les paramètres du modèle obtenu sont donnés au tableau 3.3 où d, g et s signifient déclaration, gestion et sortant. On peut constater que l'ouverture est l'événement qui excite le plus le processus de gestion ce qui n'est pas forcément surprenant quand on sait que les appels de gestions sont plus fréquents sur le début de la gestion d'un dossier et que cet événement ne peut se produire qu'une fois. Ensuite, on peut constater que ce sont les appels de gestion eux-mêmes qui se génèrent le plus ce qui va dans le sens des observations précédentes mais laisse toujours perplexes quant à l'utilité de ces appels répétés. Autre fait notable, l'envoi d'un pli ou l'ouverture d'un ordre de mission semble désexciter l'arrivée d'un appel de gestion. Une interprétation possible de ce dernier constat est que dans le cadre d'une mission le dossier va être entre les mains d'un autre professionnel (l'expert ou le réparateur) le temps de la mission et ne devrait donc pas donner lieu immédiatement à un contact avec le gestionnaire et que dans le cadre du pli envoyé on peut limiter les appels ayant uniquement pour but de connaître l'avancement du dossier. En ce qui concerne les appels sortants on pourrait s'attendre à observer un phénomène identique que pour les plis envoyés mais dans la pratique le client n'est pas forcément disponible au moment de l'appel et le simple fait d'avoir un appel manqué de son assureur peut le motiver à le rappeler.

β	μ	α_j						
		ouverture	appel d	appel g	appel s	mission	pli envoyé	pli reçu
0,04	-3,21	1,27	0,42	0,68	0,04	-0,05	-0,14	0,21

TABLE 3.3 : Estimation des paramètres du modèle d'intensité pour les appels de gestion

Il est à noter que sur l'ensemble de test on obtient une AUC de 0,710 et que ce modèle d'intensité d'appel permet de détecter 165 appels à capacité donnée de 500. Ces performances sont donc inférieures à celles obtenues par les modèles du chapitre 2 mais ne se basent uniquement que sur les dates d'événements et proviennent d'un modèle clairement interprétable. Si l'on décide de s'affranchir de la constance des β_j on se retrouve à devoir calibrer une fonction à 7 paramètres alors qu'un simple appel à cette fonction prend environ une minute. Il paraît donc peu raisonnable d'employer des techniques d'optimisation coûteuses en appels de fonction on évitera donc les méthodes utilisant une approximation du gradient. Une recherche sur une grille de valeurs possibles semble également être peu pertinente car pour tester 4 valeurs pour chaque paramètre il faudrait déjà $7^4 = 2401$ minutes. La valeur constante de $\beta = 0,04$ peut être un bon point de départ et on peut peut calibrer successivement les paramètres β_j par dichotomie si l'on considère que les effets sont séparables.

En utilisant ce principe on trouve les candidats pour les paramètres du modèle initialement posé à l'équation (3.3) dans le tableau 3.4. On peut constater une cohérence globale avec la précédente estimation tout en observant que les effets des appels entrants (de déclaration ou de gestion) semblent avoir une durée de vie supérieure aux autres événements. Sur l'ensemble de test on obtient une AUC de 0,728 et on détecte 170 appels à capacité donnée de 500. On peut considérer ce modèle comme un indicateur de la chaleur de l'activité d'un dossier. Il est remarquable de constater que l'occurrence de tous les événements augmente la probabilité d'appel à l'exception des missions et des plis envoyés ce qui suggère que l'occurrence de ces deux événements inhibe l'envie d'appeler de l'assuré.

	ouverture	appel d	appel g	appel s	mission	pli envoyé	pli reçu	μ
α_j	1.67	0.45	0.53	0.12	-0.23	-0.15	0.34	-3.15
β_j	0.12	0.02	0.02	0.05	0.15	0.07	0.10	

TABLE 3.4 : Estimation des paramètres du modèle d'intensité avec des β_j variables

3.4.3 Utilisation comme variable prédictive

La disponibilité d'un indicateur de la chaleur actuelle d'un dossier peut également servir de variable prédictive au sein d'un modèle. Si l'on reprend les modèles du chapitre précédent rajouter le score de chaleur construit permet de gagner entre 0,001 et 0,003 sur l'AUC et entre 1 et 4 appels détectés à capacité donnée de 500. L'apport sur les performances est donc minime mais l'intérêt de ce score de chaleur est de résumer les informations concernant les événements passés et on peut donc avoir bon espoir de proposer un modèle plus simple en l'utilisant tout en conservant des performances similaires à celles des modèles complexes évalués.

Dans le but de proposer un modèle très facilement explicable on fait donc le choix d'évaluer une régression logistique qui ne prend en compte plus que 6 variables à savoir

- le score de chaleur calculé précédemment,
- le mode de gestion,

- la typologie de sinistre,
- le mode de contact à la déclaration,
- la situation en intérêt-client ou non,
- la nature du dernier événement.

On s'est autorisé la connaissance du dernier événement d'une part car cela permet d'identifier une éventuelle répétition des appels et d'autre part car c'est une information facilement explicable au gestionnaire. La courbe ROC ainsi que le graphique de cohérence de ce modèle simplifié sont donnés dans la figure 3.8, il permet d'obtenir une AUC de 0,749 et de détecter 189 appels à capacité donnée de 500. On approche donc de très près les performances de la régression logistique du précédent chapitre et on reste assez compétitifs face aux autres modèles en ayant un modèle très simple et parfaitement explicable.

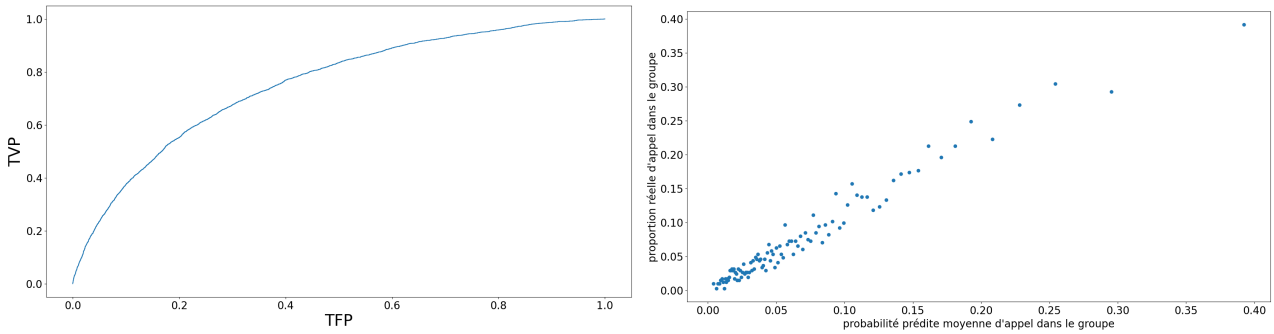


FIGURE 3.8 : Courbe ROC et cohérence du modèle simplifié

Dans une situation où l'on devra expliquer les résultats du modèle, et notamment dans le cas d'une utilisation par le gestionnaire on pourra préférer cette alternative.

3.5 Travail sur les séquences d'événements

Dans cette section on se propose d'aborder la gestion d'un dossier selon une approche plus globale en donnant plus de sens à la succession des événements en les regroupant de manière à construire des séquences cohérentes.

3.5.1 Motivations

Une des raisons qui rendent difficile la modélisation de la gestion d'un dossier peut être le fait que travailler à l'échelle des événements soit une approche trop détaillée et ne permette pas de bien capter des comportements plus globaux. En effet, on peut imaginer que, plus que les événements, ce soient des séquences d'événements qui caractérisent l'activité de gestion. Pour représenter l'évolution générale de la gestion d'un dossier il peut effectivement être intéressant d'identifier des tâches qui peuvent se composer de plusieurs événements et peuvent potentiellement mieux identifier des moments de la vie d'un dossier. La demande d'une facture, son envoi par l'assuré, l'accusé de réception de la part du gestionnaire et la notification à l'assuré d'un paiement par virement peuvent aisément être considérés comme une séquence d'événements fortement liés entre eux et afférents à une même tâche.

On va donc chercher à construire des séquences les plus cohérentes possibles avec les informations dont on dispose sur les événements. Ces séquences devront dans l'idéal permettre de représenter la gestion d'un dossier en quelques séquences, apporter des informations sur la survenance des appels de gestion ou encore aider à modéliser la gestion d'un dossier.

3.5.2 Solutions étudiées

On se propose de construire et d'utiliser deux approches assez simples et intuitives pour construire les séquences. L'une sera basée sur une approche temporelle et l'autre essaiera de capter l'objet des interactions.

Construction des séquences par temps d'inactivité

Dans la mesure où les événements possèdent un ordre d'arrivée mais aussi une date on peut mettre en place une approche basée sur la proximité temporelle pour opérer les regroupements d'événements. L'idée est ici de considérer que dans la mesure où deux événements arrivent à des dates proches c'est qu'ils concernent probablement la même tâche, le même moment de la vie du dossier. Le principe va donc être d'agréger les événements en séquences en créant des liens d'une certaine durée comme spécifié dans la définition suivante.

Définition 3.5 (Séquences par temps d'inactivité maximum)**.** Soient $\left(e_i^d, t_i^d\right)_{i\in\{1,\dots k_d\}}$ l'ensemble des couples $(\textit{événement}, \textit{date})$ du dossier d classés du plus ancien au plus récent avec k_d le nombre total d'événements du dossier d. On affecte un numéro de séquence s_i basé sur le temps d'inactivité $TI > 0$ au i-ème évènement de manière récurrente avec

$$s_1 = 0,$$
$$s_{n+1} = \begin{cases} s_n & \text{si } t_{n+1} - t_n < TI, \\ s_n + 1 & \text{sinon.} \end{cases}$$

Selon cette approche l'affectation d'un événement à une séquence est complètement déterminée par la valeur de TI. On cherche une durée maximale de lien TI qui permettent de réaliser le meilleur arbitrage possible entre agrégation d'événements ayant un faible lien entre eux et séparation d'événements probablement liés. Naturellement cet arbitrage revient à effectuer un choix sur le nombre de séquences au sein d'un dossier. La figure 3.9 donne le nombre moyen de séquences par dossier en fonction de la valeur de TI. On peut y voir que dès que l'on s'autorise un cours laps de temps pour relier des événements entre eux on parvient à diminuer sensiblement le nombre de séquences par dossier et on peut être assez confiant sur le fait que deux événements séparés de moins de 24 heures aient de fortes chances de concerner un même sujet.

Dans la pratique on peut rapprocher le choix de TI à celui du nombre de groupes dans un partitionnement en K-moyennes. Si on applique une analyse graphique avec une méthode du coude on identifie que le gain en termes de réduction du nombre de séquences devient quasi linéaire à partir d'un temps d'inactivité de deux semaines. Ce délai de 14 jours semble assez cohérent d'un point de vue pratique et permet de travailler avec un nombre plus restreint de séquence. En fixant $TI = 14$ la figure 3.10 donne la répartition du nombre de séquences par dossier qui s'établit en moyenne à 2,22. On constate une disparité du nombre de séquences avec 38,8% des dossiers qui seront clos sans qu'il ne se soit jamais écoulé plus de deux semaines sans événements mais certains dossiers qui peuvent contenir un grand nombre de séquences. Cette disparité traduit bien les écarts de complexité de gestion d'un dossier à l'autre.

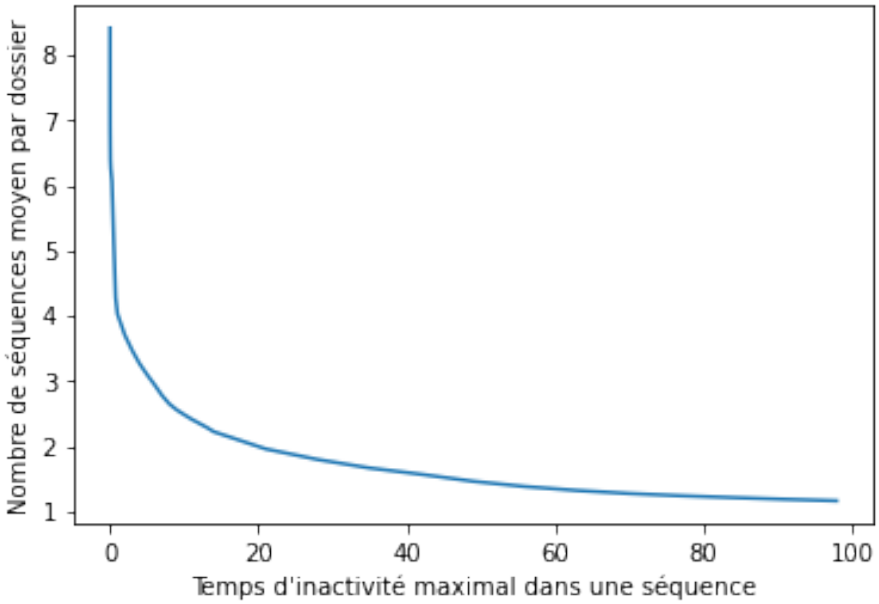


FIGURE 3.9 : Nombre moyen de séquences par dossier en fonction de la valeur de TI en jours

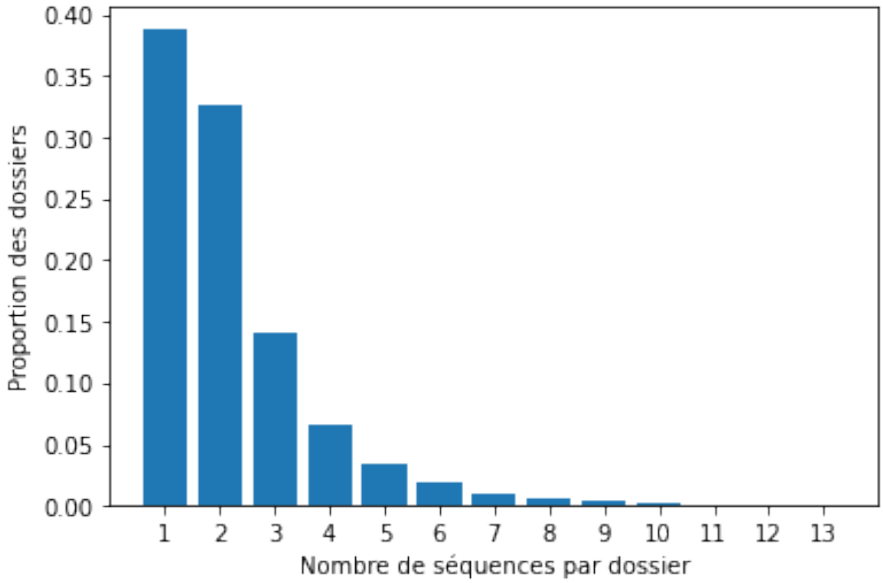


FIGURE 3.10 : Répartition du nombre de séquences par dossier avec $TI = 14$

Construction des séquences grâce à la typologie des plis

La seconde méthodologie qui sera abordée se basera sur la typologie des plis. En effet, pour chaque pli on dispose d'une typologie (un objet) qui est renseignée par le gestionnaire que le pli ait été reçu ou envoyé. On sait que cette typologie est pré-remplie automatiquement et que l'exactitude du renseignement de cette information est loin d'être parfaite. Cependant, cette typologie de pli est la seule information sur la nature des échanges que l'on a à disposition à grande échelle et mérite donc de s'y intéresser si l'on veut travailler sur le sens des échanges. L'idée utilisée sera la suivante, on peut

raisonnablement penser que lors d'une étape importante de la gestion d'un dossier au moins un pli sera envoyé ou reçu et que la typologie de ces plis devrait être une bonne information quant à la nature des échanges et des tâches qui se déroulent à proximité de la réception ou de l'envoie de ces plis.

Typologies des plis reçus Au total on dispose de 78 typologies différentes pour les plis reçus dont on donne la répartition des 15 les plus représentées au tableau 3.5. On constate donc qu'avec 15 typologies différentes on couvre 97% des plis reçus et que certaines typologies sont assez vagues, comme "Courrier client" qui est la plus représentée, mais que d'autres sont assez claires sur la nature des échanges comme "Constat DDE" ou "Devis".

typologie	proportion	cumul
Courrier client	32,0%	32,0%
Facture	9,7%	41,6%
Devis	9,7%	51,3%
Internet non signée	9,4%	60,7%
Courrier compagnie	8,6%	69,3%
Courrier interne	8,5%	77,8%
Autre	4,1%	81,9%
Constat DDE	3,7%	85,6%
Photos	2,5%	88,1%
Courrier expert	2,0%	90,1%
Rapport d'expertise	1,8%	91,9%
PV, dépôt de plainte	1,5%	93,3%
Courrier assisteur	1,4%	94,8%
TIERS	1,2%	96,0%
Courrier adversaire	1,0%	97,0%

TABLE 3.5 : Proportion et cumul des 15 typologie de plis reçus les plus représentées

Typologies des plis envoyés Au total on dispose de 279 typologies différentes pour les plis envoyés dont on donne la répartition des 15 les plus représentées au tableau 3.6. On constate donc qu'avec 15 typologies différentes on ne couvre plus que 82% des plis reçus et qu'on est toujours en présence de typologies assez vagues ("Divers", "Saisie Libre", etc.) avec d'autres assez précises ("Règlement", "AR DECLA", etc.).

On dispose donc d'un qualificatif plus ou moins pertinent sur le contenu des plis. Pour construire des séquences autour de cette information une approche naïve pourrait être d'affecter chaque événement au pli le plus proche en termes de date et d'identifier les séquences construites par la typologie des plis. Cette approche serait assez limitante dans la mesure où l'on construirait une séquence autour de chaque pli alors qu'il y a de fortes chances que deux plis (envoyés ou reçus) puissent être en lien. On peut par exemple concevoir que les plis de constat de dégâts des eaux et ceux d'accusé de réception de constat de dégâts des eaux soient fortement liés.

Afin d'éviter un découpage trop fin des séquences et pour mieux capter la nature des échanges il parait essentiel de pouvoir regrouper certaines typologies de plis. D'une part ces regroupements devront permettre de prendre en compte le fait que certains plis sont clairement des réponses à d'autres plis, mais ils devront également permettre de rapprocher des plis qui semblent être similaires. On va donc devoir regrouper certaines typologies de plis pour former des catégories de plis basées sur le contenu (et pas sur l'émetteur).

typologie	proportion	cumul
Règlement Virement	19,7%	19,7%
Mission expert/Réparateur à contacter	12,2%	32,0%
XSOC-DIVERS	8,2%	40,2%
ADVR-RECOURS - ADV/ASS	7,0%	47,2%
INFO APPEL CLIENT	6,0%	53,2%
RDDE-AR DDE	4,8%	57,9%
AR DECLA MAAF WEBHELP/CMR	4,7%	62,6%
IRSI - CLIENT AU 01.07.20	3,7%	66,2%
Saisie Libre	3,2%	69,4%
Règlement chèque	3,1%	72,5%
RDRO-RECOURS	2,8%	75,3%
AR DECLA BPCE WEBHELP/CMR	2,0%	77,2%
PSOC-PROPOSITION	1,8%	79,0%
RRCF-AR RCF	1,6%	80,6%
Règlement Virement Lien internet	1,5%	82,1%

TABLE 3.6 : Proportion et cumul des 15 typologie de plis envoyés les plus représentées

Pour réaliser les groupes de typologies on adopte la stratégie suivante. On commence par représenter les typologies par des points dans $\mathbb{R}^n, n \in \mathbb{N}$ de sorte à ce que soient positionnées proches les typologies intervenant dans des situations proches. Par la suite on pourra appliquer des techniques de clustering pour constituer des groupes cohérents.

Afin de représenter les typologies de plis dans un espace permettant de constituer les groupes on utilisera le principe suivant. On représente tous les événements d'un dossier sous la forme d'une phrase où les mots sont les événements en prenant la typologie de pli pour les événements qui sont des plis. On dispose donc d'une phrase par dossier qui décrit l'évolution d'un dossier en conservant l'ordre chronologique des événements mais sans avoir la précision temporelle sur les dates d'occurrences. Pour un dossier la phrase de ses événement peut donc par exemple être

ouverture appel_de_déclaration constat_dde appel_de_gestion ar_constat_dde appel_sortant règlement_virement clôture.

On va donc chercher à représenter un événement (un mot) par une valeur qui permette d'identifier son contexte pour pouvoir rapprocher les mots ayant des contextes similaires. Pour ce faire on utilisera l'algorithme de Word2vec, MIKOLOV et al. (2013), employé pour représenter par un vecteur numérique le contexte d'un mot. On utilisera ici la version CBOW (Continuous Bag Of Words) de l'algorithme Word2vec. Le principe est de d'entraîner un modèle qui selon les mots voisins (selon une fenêtre de taille définie) d'un mot cible prédit ce mot cible. Si on fixe une fenêtre qui ne prend en compte que le mot précédent et le mot suivant et qu'on reprend l'exemple de "phrase" ci-dessus le modèle cherchera à prédire règlement_virement s'il sait que les deux mots de la fenêtre d'observation sont appel_sortant et clôture. Le modèle utilisé dans l'algorithme est un réseau de neurones ayant une couche cachée et prenant en entrée la représentation en encodage one-hot des mots présents dans la fenêtre (le contexte du mot à prédire) et ayant en couche de sortie un vecteur qui donne la probabilité du mot cible sous le format d'un encodage one-hot.

La figure 3.11 donne la structure du réseau utilisé pour le CBOW. Il est à noter que la matrice des poids des liens entre la couche d'entrée et la couche cachée est commune à chaque position de mot ce

qui signifie que la position d'un mot dans la fenêtre n'impacte pas la prédiction du mot cible, seule compte sa présence dans le contexte du mot cible. Lorsqu'on multiplie la représentation en encodage one-hot d'un mot par cette matrice de poids on obtient donc la représentation contextualisée du mot dans le sens ou cette représentation est utile pour prédire la présence du mot cible. Des mots présents dans des contextes similaires devraient donc avoir une représentation contextualisée proche.

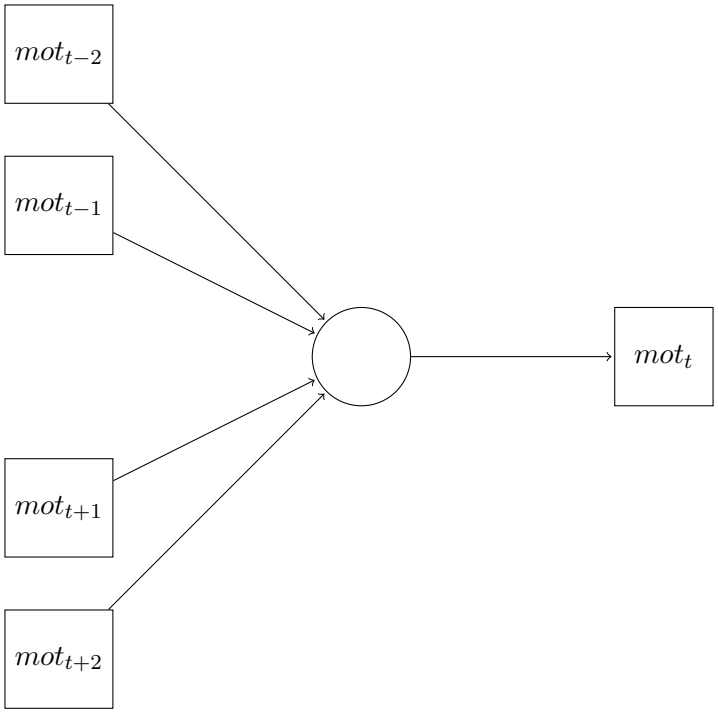


FIGURE 3.11 : Structure du réseau de neurones utilisé dans la version CBOW de l'algorithme de Word2vec

L'entraînement du réseau dans la tâche de prédiction du mot cible permet d'ajuster les poids et l'amélioration des performances doit logiquement passer par des poids qui permettent de mieux représenter le contexte d'un mot. Si n est la dimension de la couche cachée, on est donc en présence d'une représentation dans $\mathbb{R}^n$ des mots (ou événements) utilisés pour produire les "phrases" décrivant la gestion des dossiers.

Une fois réalisée la représentation contextualisée des mots on ne conserve que les mots qui correspondent à des plis, donc des typologies de plis. On va à présent pouvoir réaliser un regroupement sur les typologies de plis. Pour cela on va utiliser un algorithme de classification ascendante hiérarchique avec lien moyen et la similarité cosinus qui est généralement utilisée pour mesurer la proximité de la représentation vectorielle de deux mots après un Word2vec. L'algorithme de classification ascendante hiérarchique part donc avec autant de classes qu'il y a de typologies et à chaque itération il rassemble les deux classes A et B qui maximisent la quantité suivante

$$\frac{1}{\#A \times \#B} \sum_{t_A \in A} \sum_{t_B \in B} \cos \theta_{t_A, t_B},$$

où θ_{t_A,t_B} est l'angle formé par les représentation dans $\mathbb{R}^n$ des typologies de plis t_A et t_b obtenues précédemment. Dans la suite on va utiliser une représentation en $n = 5$ dimensions sur une fenêtre qui prend en compte les 3 mots précédents et les 3 mots suivants pour le Word2vec. On construira ensuite 20 classes de typologies de plis par classification ascendante hiérarchique. Si on ordonne les classes

par rapport au volume global de plis qu'elles vont contenir de manière décroissante on peut obtenir la répartition des occurrences de chaque classe que l'on donne au tableau3.7 dans lequel on précise le nombre de typologies de plis différentes présentes dans la classe. On couvre donc naturellement tous les plis avec 20 classes. Fait assez remarquable, la classe 8 qui est la classe regroupant (de loin) le plus de typologies de plis différentes ne représente que moins de 1% du volume total de plis. On peut donc regrouper les classes 8 et plus en une seule classe sans perdre trop d'information.

classe	proportion	cumul	typologies
1	31,1%	31,1%	12
2	24,3%	55,4%	12
3	12,2%	67,6%	34
4	10,9%	78,5%	12
5	8,1%	86,6%	8
6	8,1%	94,7%	8
7	2,9%	97,5%	19
8	0,8%	98,3%	200
9	0,7%	98,9%	18
10	0,4%	99,4%	4
11	0,3%	99,7%	5
12	0,2%	99,8%	1
13	0,1%	99,9%	2
14	0,1%	100,0%	6
15	0,0%	100,0%	3
16	0,0%	100,0%	3
17	0,0%	100,0%	3
18	0,0%	100,0%	2
19	0,0%	100,0%	3
20	0,0%	100,0%	2

TABLE 3.7 : Proportion et cumul des 20 classes de typologie de plis envoyés ou reçus

On a donc regroupé les plis (reçus et envoyés) en 8 classes dont on espère qu'elles segmentent bien le type de tâche en cours. On peut se rassurer sur ce point en observant par exemple que la deuxième classe concerne des typologies relatives à la demande de pièces justificatives, l'envoi de devis, de factures ou de photos et la notification de règlement par différents moyens. On constate également que la septième classe concerne des échanges mettant en jeu une partie adverse ou encore que la quatrième classe semble regrouper les plis concernant la mise en relation avec un expert ou un réparateur tandis que la troisième regroupe les échanges relatifs à des constats amiables. Il semblerait que la première classe regroupe l'ensemble des typologies de plis génériques comme "courrier client" par exemple. Enfin on se sert de ces 8 classes de typologies pour définir les séquences d'événements de la manière suivante.

Définition 3.6 (Séquences par typologie de pli)**.** Soit d un dossier pur lequel au moins un pli a été envoyé ou reçu. On va affecter à chaque événement (pas uniquement les plis) une des classes construites sur les typologies de plis de la manière suivante. La classe du i-ème événement $c(i)$ est définie de la façon suivante

$$c(i) := \text{la classe du pli le plus proche du i-ème événement en terme de dates.}$$

La classe d'un pli définie de cette façon correspond bien à celle obtenue précédemment. On affecte un numéro de séquence s'_i basé sur les classes des événements au i-ème évènement de manière récurrente avec

$$s'_1 = 0,$$
$$s'_{n+1} = \begin{cases} s'_n & \text{si } c(i+1) = c(i), \\ s'_n + 1 & \text{sinon.} \end{cases}$$

La stratégie est donc la suivante, on considère que le sujet d'un événement est le même que celui du pli le plus proche et tant que le sujet ne change pas on considère que la séquence est en cours. Dans le cas où un dossier ne comporterait aucun pli dans sa gestion on considérera que tous les événements sont dans la première séquence. La figure 3.12 donne la répartition du nombre de séquences comme définies ci-dessus par dossier qui s'établit en moyenne à 2,94. On est donc en présence de plus de séquences par dossiers que pour les séquences définies par le temps d'inactivité et on constate à nouveau une disparité qui semble cohérente.

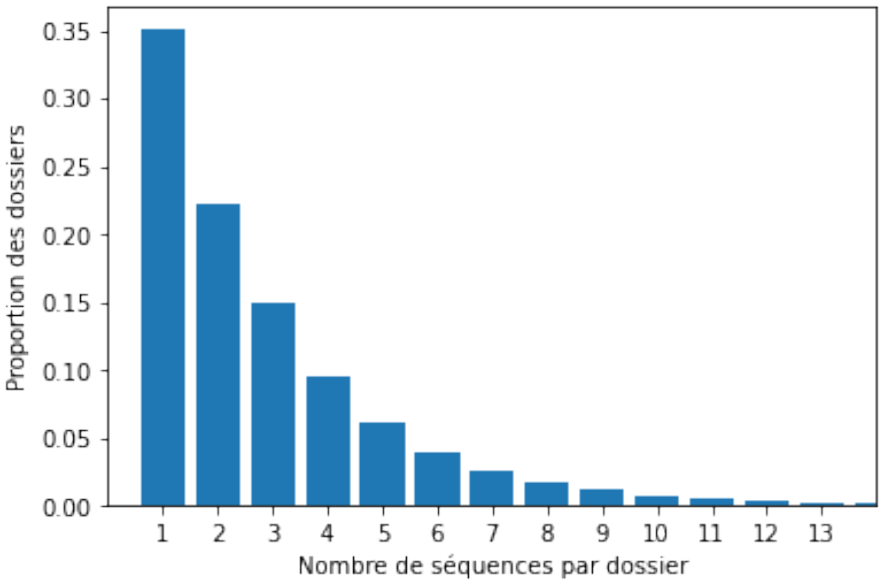


FIGURE 3.12 : Répartition du nombre de séquences basées sur les typologies de plis par dossier

3.5.3 Résultats

On se propose ici d'analyser les séquences construites selon les deux méthodologies utilisées. Pour identifier une séquence on dispose dans les deux cas de son numéro (s ou s') mais dans le second cas on préférera identifier une séquence par sa classe plutôt que son numéro pour avoir une approche plus basée sur la nature des échanges. Dans le cas d'un dossier sans pli les événements seront affectés à une classe 0.

Répartition des séquences et nombre d'appels par séquence

On s'intéresse ici à la proportion dans laquelle chaque séquence est représentée dans la gestion des dossiers mais aussi au nombre d'appels de gestion par séquence. On représente ces informations pour les deux méthodologies aux tableaux 3.8 et 3.9 en ayant restreint le premier tableau à la dixième séquence définie par temps d'inactivité. Si dans le premier cas la part que représente une séquence par son numéro est directement reliée à la répartition du nombre de séquences par dossier dans le

second cette donnée est plus propre à la nature des échanges. Sur le premier tableau on constate que les échanges semblent de moins en moins faire intervenir d'appels de gestion au fur et à mesure que les périodes d'activité se succèdent. Sur le deuxième tableau il est possible de voir que les séquences issues de la quatrième classe comportent plus d'appels de gestion que les autres en restant une séquence assez régulièrement présente. Ce constat s'applique aussi aux séquences issues de la deuxième classe qui semblent concerner le moment de la gestion d'un dossier où les pièces nécessaires à un règlement sont demandées et transmises provoquant le règlement.

numéro de séquence	proportion	appel de gestion
1	44,9%	0,436
2	27,5%	0,343
3	12,8%	0,340
4	6,5%	0,335
5	3,5%	0,336
6	2,0%	0,316
7	1,1%	0,278
8	0,7%	0,253
9	0,4%	0,218
10	0,2%	0,228

TABLE 3.8 : Proportion et nombre d'appels de gestion par séquence pour les 10 premières séquences définies par temps d'inactivité

classe de séquence	proportion	appel de gestion
0	4,0%	0,162
1	25,8%	0,313
2	22,4%	0,370
3	12,2%	0,267
4	9,7%	0,447
5	11,1%	0,131
6	8,7%	0,242
7	3,1%	0,245
8	2,9%	0,180

TABLE 3.9 : Proportion et nombre d'appels de gestion par séquence pour les séquences définies par classe

Transitions entre séquences

Afin de modéliser l'évolution de la gestion d'un dossier on peut donc avoir bon espoir que l'utilisation des transitions entre séquences permettent d'obtenir des résultats plus concluants qu'en se basant sur les transitions entre événements. On ne s'intéressera ici uniquement qu'aux séquences définies par classe dans la mesure où les séquences définies par temps d'inactivité identifiées par leurs numéros possèdent un ordre d'enchaînement déterministe avec un aléa qui porte uniquement sur le nombre de séquences. On espère donc que le processus discret formé par la succession des séquences définies par classes d'un dossier forme un chaîne de Markov. Afin de représenter cette modélisation sous la forme d'une matrice de transition on rajoute une séquence d'ouverture et de clôture en début et fin de gestion de chaque dossier. L'état d'ouverture sera donc l'état initial de la chaîne et permettra d'observer

la loi de distribution de la première séquence tandis que l'état de clôture sera un état absorbant et permettra d'observer les séquences permettant de terminer le processus de gestion.

On donne la matrice de transition ainsi obtenue au tableau 3.10 ainsi qu'une représentation plus visuelle à la figure 3.13. On peut observer assez logiquement que c'est la séquence de classe 2 qui est la plus souvent la dernière étant donné qu'elle est liée à l'envoi de justificatif et au règlement. On peut également voir que pour les séquences de classe 4 (qui concernent la mise en relation avec un expert ou un réparateur) dans près d'un tiers des cas la prochaine séquence sera de classe 2. De manière générale on ne constate pas d'aberration évidente. Les séquences de classe 0 représentent l'ensemble des événements pour les dossiers n'ayant aucun pli et une séquence de cette classe suit forcément l'ouverture et précède forcément la clôture. Ainsi les lignes associées à la classe 0 et à la clôture sont égales. L'utilité de la classe 0 est de ne pas laisser des événements non affectés à une séquence.

	ouv	0	1	2	3	4	5	6	7	8	clô
ouv	0,000	0,115	0,191	0,129	0,085	0,068	0,301	0,057	0,032	0,023	0,000
0	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000
1	0,000	0,000	0,000	0,289	0,120	0,089	0,009	0,103	0,019	0,033	0,338
2	0,000	0,000	0,239	0,000	0,054	0,088	0,002	0,051	0,011	0,015	0,541
3	0,000	0,000	0,330	0,190	0,000	0,097	0,009	0,073	0,013	0,017	0,271
4	0,000	0,000	0,253	0,324	0,110	0,000	0,004	0,090	0,023	0,016	0,181
5	0,000	0,000	0,256	0,169	0,225	0,066	0,000	0,074	0,058	0,031	0,121
6	0,000	0,000	0,301	0,213	0,109	0,079	0,008	0,000	0,021	0,020	0,249
7	0,000	0,000	0,242	0,215	0,058	0,080	0,016	0,078	0,000	0,022	0,288
8	0,000	0,000	0,381	0,206	0,072	0,085	0,008	0,059	0,018	0,000	0,172
clô	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000

TABLE 3.10 : Matrice de transition pour la succession des séquences construites par classe

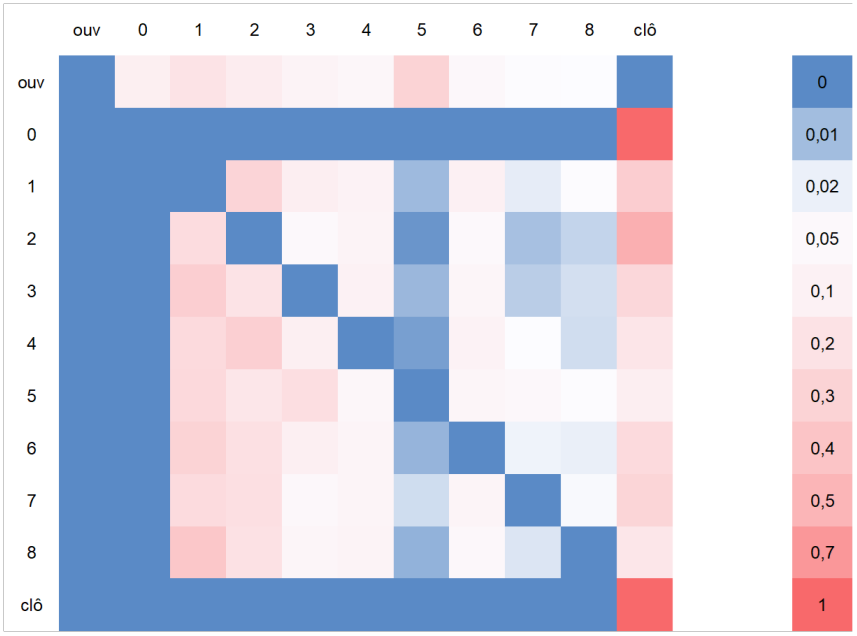


FIGURE 3.13 : Représentation graphique de la matrice de transition pour la succession des séquences construites par classe

Afin de juger de la capacité de cette matrice de transition à représenter la succession des séquences au cours de la gestion d'un dossier on se propose de regarder si des simulations qui en sont issues sont semblables aux observations réelles. Étant donné que l'on cherche à déterminer si un processus discret suit bien une chaîne de Markov on pourrait également procéder à un test statistique sur l'indépendance du passé au-delà du dernier état en comparant la matrice de transition avec des matrices de transition d'ordres supérieurs. Cependant, comme la matrice de transition comporte un état absorbant et des coefficients nuls ou presque nuls on préférera étudier la loi de l'enchaînement global des séquences au sein d'un dossier. Pour ce faire on note $T(d)$ la trajectoire complète d'un dossier (en tenant compte de l'ordre d'arrivée des séquences) et des valeurs possibles de $T(d)$ sont donc {0}, {5,1,2} ou encore {1,4,1,5,2} par exemple. On veut donc observer si la matrice de transition obtenue conduit à des trajectoires cohérentes avec celles observées. On représente les proportions réelles observées de trajectoires ainsi que celles obtenues avec un million de simulations utilisant la matrice de transition au tableau 3.11 en se limitant au 15 trajectoires les plus fréquentes sur les observations réelles.

Trajectoire	proportion réelle	proportion estimée
0	11,48%	11,55%
2	7,50%	6,97%
1	6,67%	6,45%
5	3,61%	3,64%
3	2,84%	2,29%
5 2	2,75%	2,76%
5 1	2,67%	2,59%
1 2	2,63%	2,94%
5 3	1,79%	1,82%
6	1,51%	1,41%
4	1,46%	1,24%
5 1 2	1,30%	1,22%
4 2	1,14%	1,20%
7	1,04%	0,92%
2 1	0,95%	1,00%

TABLE 3.11 : Proportion des trajectoires réelles et simulées

On constate une forte similarité de la distribution des trajectoires entre données réelles et simulations et l'on peut mesurer l'écart de la façon suivante.

Définition 3.7 (Écart de distribution)**.** Soit $\mathcal{T}$ l'ensemble des trajectoires obtenues au moins une fois dans les données réelles ou par simulation, pour $T \in \mathcal{T}$ on note $p(T)$ et $\hat{p}(t)$ les proportions réelle et estimée de la trajectoire T et on définit $\mathcal{E}$ l'écart de distribution par

$$\mathcal{E} := \frac{1}{2} \sum_{T \in \mathcal{T}} |p(T) - \hat{p}(T)|.$$

Si on considère p et $\hat{p}$ comme des lois de probabilité la définition correspond à celle de la distance en variation totale entre deux mesures de probabilité. On obtient un écart de distribution de $\mathcal{E} = 0,102$ ce qui est assez satisfaisant car cela signifie qu'en utilisant la matrice de transition la distribution des trajectoires a 90% de sa masse qui correspond aux observations réelles. Cette observation est assez remarquable dans la mesure où l'on s'intéresse à l'exactitude des trajectoires. On peut également espérer que la répartition des séquences simulées soit cohérente, on donne la comparaison au tableau 3.12 et l'on observe bien une très forte similarité. En effectuant un test du χ^2 sur l'hypothèse nulle de l'égalité des proportions on obtient un $p - valeur$ de 0,833 et on accepte donc cette hypothèse.

On peut donc être assez confiant sur la capacité d'une chaîne de Markov à simuler des trajectoires de gestion cohérentes en se basant sur les séquences définies par classe.

classe de séquence	proportion réelle	proportion estimée
0	4,00%	4,03%
1	25,82%	25,81%
2	22,42%	22,38%
3	12,19%	12,19%
4	9,75%	9,74%
5	11,08%	11,10%
6	8,75%	8,73%
7	3,09%	3,11%
8	2,91%	2,92%

TABLE 3.12 : Proportion des séquences définies par classe sur trajectoires réelles et simulées

Ces séquences définies par classe permettent de poser un cadre général sur le processus de gestion que l'on a bien pu modéliser. Pour aller plus loin et disposer d'un modèle plus global il faudrait également modéliser le temps qui sépare les séquences, la durée des séquences ainsi que la répartition des événements au sein des séquences.

Il est à noter que repérer et définir des séquences d'événements est une problématique assez répandue et peut faire l'objet de techniques plus poussées que celles évoquées ici. On peut notamment évoquer l'utilisation d'arbres des suffixes qui peuvent être utiles pour détecter des motifs et construire des algorithmes permettant de détecter des séquences similaires mais non forcément égales.

3.6 Conclusion

Ce chapitre aura permis de mettre en avant des pistes de modélisation globale du phénomène de gestion ainsi que certains résultats théoriques notamment sur la génération des appels de gestion. En effet, si ce chapitre a pu révéler la difficulté de mettre en place une modélisation globale il aura été utile pour mettre en avant une capacité théorique de la communication écrite sortante pour limiter le nombre d'appels de gestion. Dans une perspective d'amélioration de l'identification des dossiers à risques d'appels on pourrait également utiliser les séquences construites pour établir un indicateur d'écart de nombre d'appels entre le passé d'un dossier et une réalisation moyenne selon les séquences parcourues. Un tel indicateur pourrait renseigner sur la propension à appeler pour un dossier à tâches équivalentes.

Chapitre 4

Utilisations des résultats de l'étude

Les différents résultats de l'étude conduite au cours de ce mémoire peuvent être utiles dans la compréhension des interactions entre l'assuré et l'assureur au cours de la gestion et de l'indemnisation d'un sinistre. L'objet principal de l'étude étant de comprendre la survenance des appels de gestion dans le but de mieux les maîtriser voire de les limiter, il sera intéressant de voir quelles perspectives offrent les quelques résultats de l'étude.

4.1 Analyse des dossiers à haut risque d'appel

La modélisation de la probabilité d'appel ayant conduit à des résultats assez satisfaisants et cohérents, il est possible d'aller analyser le comportement futur des dossiers en fonction de la prédiction pour identifier les pratiques qui pourraient être vertueuses.

4.1.1 Appels et autres événements futurs des dossiers à risque

Dans la suite on affecte une probabilité d'appel à chaque dossier en considérant la moyenne des modèles de forêt aléatoire, de XGBoost et de réseau de neurones présentée en section 2.6.2. On considère de plus qu'un dossier a une forte probabilité d'appel si la probabilité d'appel obtenue est supérieure à 0,2. Dans la pratique ces dossiers à forte probabilité d'appel représentent 6,8% du portefeuille des dossiers ouverts et parmi ces dossiers 29,6% vont effectivement générer au moins un appel à deux semaines. Le tableau 4.1 donne la moyenne et l'écart-type du nombre de chaque événement des dossiers à forte probabilité d'appel dans les deux semaines. Il est possible d'y voir que si 29,6% des dossiers génèrent au moins un appel il y a 0,527 appel entrant par dossier ce qui signifie que les dossiers qui appellent génèrent en moyenne $\frac{0,527}{0,296} = 1,78$ appels. Dans 17,9% des cas le dossier sera clôturé dans les deux semaines et dans l'ensemble on peut assez logiquement constater que les dossiers ayant une forte probabilité d'appels ont aussi pour les autres événements une activité marquée.

événement	moyenne	écart-type
appel entrant	0,527	1,121
appel sortant	0,291	0,687
mission	0,098	0,301
pli envoyé	0,539	0,791
pli reçu	0,549	1,040
clôture	0,179	0,383

TABLE 4.1 : Moyenne et écart-type du nombre de chaque événement des dossiers à forte probabilité d'appel dans les deux semaines

4.1.2 Actions à disposition du gestionnaire et politique de Meilleure Prochaine Action générale

Dans le but de réduire la charge d'appels de gestion la possibilité d'identifier les dossiers à risque doit s'associer à des mesures permettant de les inhiber. En effet, l'idée de proactivité est séduisante car si cette dernière est efficace elle pourrait permettre de limiter le nombre d'appels entrants tout en étant généralement bien perçue du point de vue de la satisfaction client. La probabilité d'appel peut servir de score de priorisation tandis que les résultats du tableau 3.4 suggèrent que au-delà de la décision de créer un ordre de mission qui est un choix ayant un impact important sur le coût de gestion seul l'envoi d'un pli semble en mesure de freiner l'appel.

Afin de mesurer l'impact de la prise de décision du gestionnaire on se propose de segmenter les dossiers à forte probabilité d'appel en fonction de l'action qui suit la date à laquelle a été évaluée cette probabilité. La segmentation se fera selon les possibilités suivantes.

- Aucune action n'est arrivée dans les deux semaines ou la prochaine action était un appel entrant.
- La première action dans les deux semaines était un pli envoyé.
- La première action dans les deux semaines était un appel sortant.
- La première action dans les deux semaines était un pli reçu.

On a donc volontairement omis la possibilité de l'ouverture d'un ordre de mission car cet événement bien qu'en partie à la main du gestionnaire représente un coût important et vient modifier profondément la suite du processus de gestion et d'indemnisation. On a également inclus la réception d'un pli bien que cette action ne soit pas à l'initiative du gestionnaire d'une part pour analyser son impact mais également pour aller étudier l'éventuel impact du délai de traitement de ces plis reçus pour les dossiers à forte probabilité d'appel.

Dans l'optique de mesurer l'impact de ces actions on modifie légèrement la variable observée afin de ne pas biaiser l'analyse. On se propose de regarder si un appel de gestion est intervenu dans la fenêtre des deux semaines suivant l'arrivée de cette prochaine action si elle existe ou dans les deux semaines suivant l'évaluation de la probabilité s'il n'y a aucune action dans les deux semaines ou que le prochaine action est un appel entrant.

Selon ce principe d'analyse d'impact de la prochaine action on donne le parcours des dossiers à forte probabilité d'appel à la figure 4.1. On peut y voir que dans plus de 58% des cas aucune action n'intervient dans les deux semaines ou avant un appel entrant et que plus d'un tiers de ces dossiers vont occasionner au moins un appel. Dans la mesure où ces dossiers sont ceux ayant le plus fort taux d'appel cela renforce l'idée que la proactivité a de bonne chance d'être productive. On constate également que pour les dossiers à forte probabilité d'appel le conseiller n'est à l'initiative de la prochaine d'action que dans moins d'un quart des cas. Fait le plus marquant, dans le cas où la prochaine action vient du conseiller il choisit la communication écrite dans un peu plus de la moitié des cas et lorsqu'il le fait l'assuré appelle sensiblement moins alors qu'un appel sortant semble laisser neutre l'envie de l'assuré d'appeler. De plus, dans le cas où l'assuré contact le gestionnaire à l'écrit le délai de traitement du pli semble être assez neutre sur les appels de gestion. Ce constat ne signifie pas que les assurés n'appellent pas lorsqu'ils considèrent que le délai de traitement est trop long mais plus que cet effet est compensé par le fait que les échanges puissent générer des appels.

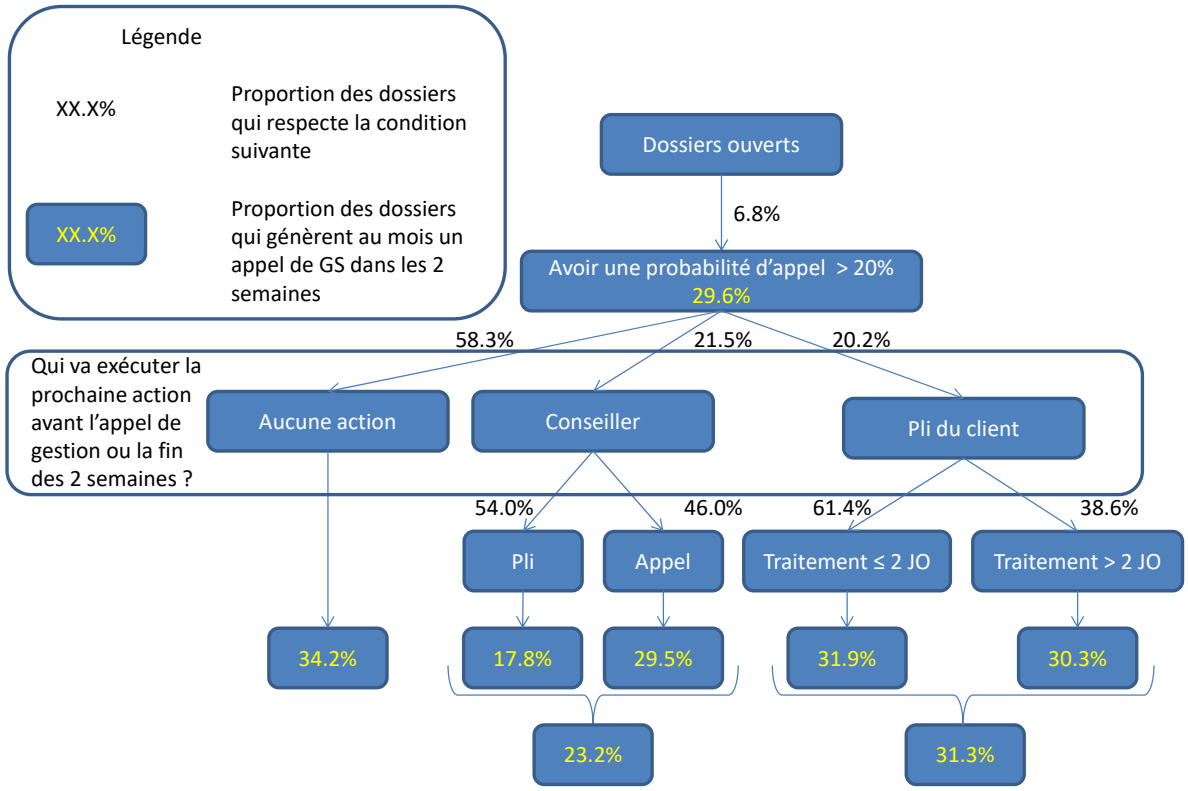


FIGURE 4.1 : Parcours des dossiers à forte probabilité d'appel

Cette analyse semble dégager un principe assez clair, la proactivité peut effectivement limiter la survenance des appels de gestion à condition que cette proactivité se fasse par une communication écrite. Une des raisons qui peut rendre l'appel sortant inefficace voire contre-productif est la problématique de l'appel non-répondu par l'assuré qui suscite un rappel mais on peut également facilement imaginer que même un appel répondu puisse donner "goût" à la communication téléphonique à l'assuré. Un des avantages de la communication écrite vers l'assuré est de s'éviter une partie des appels qui font suite à l'oubli d'une information qui aurait été donnée à l'oral.

Une première préconisation pourrait donc être de favoriser autant que faire se peut la communication sortante écrite lorsque l'on sait que l'assuré se trouve dans un moment où sa probabilité d'appel est importante.

4.1.3 Meilleure Prochaine Action par dossier par modélisation

Plus généralement il est possible de concevoir que la meilleure prochaine action par dossier dans le but de minimiser le risque d'appel soit propre à chaque dossier. En effet, il paraît peu probable que relancer par pli un dossier qui ait une faible probabilité d'appel soit pertinent mais il est également envisageable que dans certains cas un appel sortant puisse être pertinent. On s'intéressera donc ici à évaluer la probabilité d'appel conditionnée à la nature de la prochaine action. On définit donc théoriquement la meilleure prochaine action selon la définition suivante.

Définition 4.1 (Meilleure Prochaine Action)**.** Soit d un dossier ouvert à la date t, on définit la meilleure prochaine action du d en t, notée $MPA(d,t)$, comme suit

$$MPA(d,t) := \underset{a\in A}{\operatorname{argmin}}\, \mathbb{P}(AG^d_{t+t_a+14} > AG^d_{t+t_a} \mid \mathcal{F}^d_t, PA(d,t) = a),$$

avec les mêmes notations qu'à la définition 2.1 et A l'ensemble des actions possibles telles que définies ci-dessus à savoir $A = \{$aucune action, pli envoyé, appel sortant, pli reçu$\}$. On note également $PA(d,t)$ la prochaine action du dossier d après t et t_a le temps écoulé entre t et l'occurrence de cette prochaine action avec $t_{\text{aucune action}} = 0$.

On modélise la probabilité d'appel conditionnée à la prochaine action en utilisant les mêmes variables prédictives qu'au chapitre 2 auxquelles on ajoute l'information de la prochaine action. Dans le but d'obtenir une meilleure prochaine action individualisée il est nécessaire de choisir un modèle pouvant prendre en compte des interactions. On choisit donc de modéliser ces probabilités conditionnelles au moyen d'un XGBoost. Afin de se rendre compte de l'impact capté par le modèle de la prochaine action on donne au tableau 4.2 les probabilités conditionnelles pour les dossiers ayant la plus grande probabilité d'appel selon la moyenne des modèles de forêt aléatoire, de XGBoost et de réseau de neurones présentée en section 2.6.2. On constate donc que pour ces dossiers à haut risque d'appel la probabilité conditionnelle minimale est atteinte quand la prochaine action est un pli envoyé confirmant les constats précédents.

Numéro de dossier	probabilité d'appel	probabilité d'appel conditionnée à la prochaine action			
		attendre	appel sortant	pli envoyé	pli recu
1	0,785	0,840	0,685	0,633	0,754
2	0,749	0,770	0,573	0,540	0,688
3	0,737	0,791	0,526	0,508	0,685
4	0,733	0,831	0,631	0,539	0,718
5	0,722	0,769	0,539	0,445	0,681
6	0,706	0,765	0,551	0,533	0,663
7	0,700	0,710	0,541	0,452	0,602
8	0,694	0,744	0,573	0,508	0,656
9	0,693	0,664	0,576	0,490	0,606
10	0,690	0,677	0,479	0,392	0,594
11	0,687	0,748	0,530	0,435	0,609
12	0,683	0,734	0,530	0,478	0,620
13	0,682	0,777	0,581	0,532	0,660
14	0,681	0,722	0,571	0,480	0,658
15	0,681	0,647	0,502	0,366	0,525

TABLE 4.2 : probabilités d'appel conditionnelles des dossiers ayant la plus forte probabilité d'appel

Plus généralement si on identifie l'action minimisant la probabilité conditionnelle de chaque dossier on obtient la répartition de meilleure prochaine action donnée au tableau 4.3. On constate donc que presque systématiquement la meilleure prochaine action se joue entre attendre et envoyer un pli. Dans la pratique plus le dossier a une forte probabilité d'appel plus le modèle identifie l'envoi d'un pli comme étant la meilleure prochaine action ce qui peut se voir à la figure 4.2 où l'on représente la proportion de dossiers ayant pour meilleure prochaine action d'attendre ou d'envoyer un pli en fonction de la probabilité d'appel. Même une modélisation individuelle vient donc confirmer qu'une proactivité ayant pour but de limiter le nombre d'appels de gestion passe par la communication écrite. On peut également constater que le conseiller n'a pas intérêt à relancer un assuré ayant une faible probabilité d'appel.

meilleure prochaine action	proportion des dossiers
attendre	68,8%
pli envoyé	31,0%
pli reçu	0,1%
appel sortant	0,1%

TABLE 4.3 : Répartition des meilleures prochaines actions par modélisation

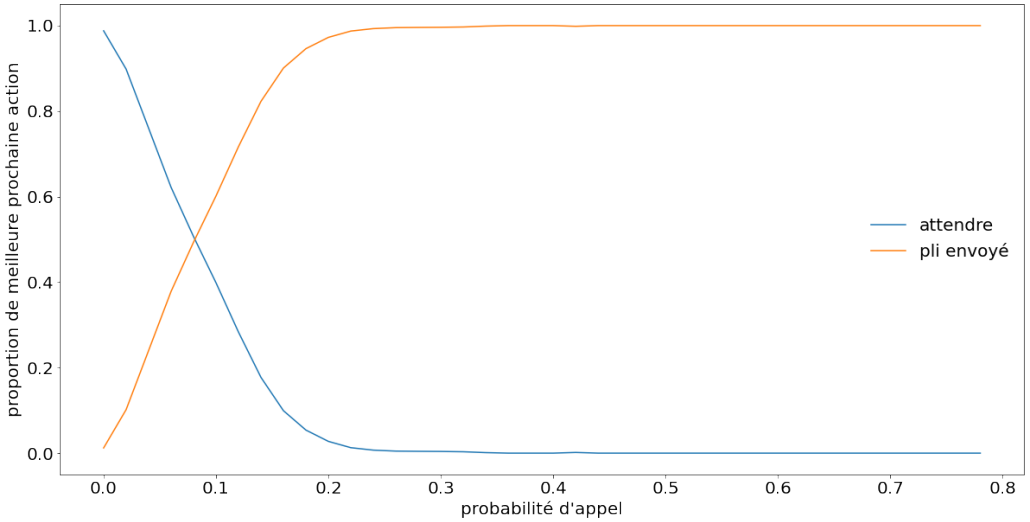


FIGURE 4.2 : Meilleure prochaine action en fonction de la probabilité d'appel

4.2 Mise en place d'un test de priorisation

La section précédente est venue renforcer l'idée qu'une proactivité (écrite) puisse avoir des effets vertueux dans la réduction de la charge d'appels de gestion. Afin d'une part de mesurer l'impact en conditions réelles de cette proactivité mais aussi de mieux en apprécier son coût il a été décidé de procéder à une expérimentation.

4.2.1 Procédure du test

La procédure du test doit permettre de tirer plusieurs conclusions. On veut premièrement s'assurer que la modélisation de la probabilité d'appel peut effectivement se calculer à date sans observer de dérive de performance. On souhaite également prendre des mesures de proactivité réelles et les analyser. Pour ce faire, le cadre d'évaluation suivant a été défini en prenant en compte des problématiques d'accès à la donnée et de disponibilité des gestionnaires. Les principes principaux seront les suivants.

1. L'évaluation de la probabilité d'appel se fait le week-end.
2. On répartit les $3n$ dossiers ayant la plus grande probabilité d'appel selon une capacité de test donnée en trois groupes de n dossiers.
   - n dossiers témoins pour lesquels aucune consigne n'est transmise et où la gestion se déroule comme à l'accoutumée.
   - n dossiers font l'objet d'un traitement automatisé avec l'envoi d'un pli standardisé spécifiant que le dossier est en cours de traitement et rappelant des informations générales notamment les canaux pour transmettre des documents.

- n dossiers font l'objet d'un traitement personnalisé rappelant les informations importantes sur l'avancement de la gestion du dossier et donnant de la visibilité sur sa suite. L'idée est de donner à l'assuré l'ensemble des informations pertinentes sur son dossier de manière synthétique.

3. On évalue si les différences de comportement dans les trois groupes sont statistiquement significatives.

4. On apprécie l'impact d'une généralisation de la proactivité en termes de réduction de la charge d'appels de gestion.

Bien que cela aurait pu être utile pour l'analyse il aurait été compliqué de constituer un groupe où l'on aurait forcé l'attente de la part du gestionnaire. La répartition des $3n$ dossiers à la plus forte probabilité d'appel se fera par tirage aléatoire. Le groupe témoin doit déjà en premier lieu permettre de se rassurer sur la modélisation. Par exemple si les $3n$ dossiers des trois groupes ont tous une probabilité d'appel supérieure à 0.2 on s'attend selon la figure 4.1 à ce que le groupe témoin ait une proportion réelle d'appel comparable à 0.296. Si le groupe témoins ne met pas en avant d'incohérence de modélisation on pourra analyser les proactivités par pli standardisé et par pli personnalisé au moyen de tests statistiques.

Test de proportion Une façon simple de se prononcer sur la présence d'un impact des deux types de proactivité est de procéder à un test de proportion en les comparant au groupe témoin. Le principe est le suivant.

Soient deux groupes A et B de n réalisations d'une variable binaire (toutes les réalisations sont indépendantes) qui dans notre cas est la présence d'appel de gestion dans les deux semaines suivant la mesure de proactivité le cas échéant ou les deux semaines suivant l'évaluation de la probabilité d'appel si le groupe est le groupe est le groupe témoin. Il est à noter que la variable binaire n'est pas exactement la même en fonction des cas mais cette approximation est acceptable surtout si le pli de proactivité est envoyé peu de temps après l'évaluation de la probabilité d'appel.

On note Π_A et Π_B les probabilités théoriques de la classe positive dans les deux groupes et on pose les hypothèse nulle et alternative, notées $\mathcal{H}_0$ et $\mathcal{H}_1$, suivantes

$$\begin{aligned}\mathcal{H}_0 &: \Pi_A = \Pi_B = \Pi,\\ \mathcal{H}_1 &: \Pi_A \neq \Pi_B.\end{aligned}$$

On note P_A et P_B les proportions observées, on considère généralement que l'on peut faire l'approximation d'une hypothèse de normalité sur P_A et P_B si

$$\begin{aligned} n &> 30,\\ \min(\Pi, 1-\Pi) &> \frac{5}{n}.\end{aligned}$$

Si ces conditions sont respectées, qu'on se place sous $\mathcal{H}_0$ et qu'on pose la statistique de test suivante

$$U := \frac{P_A - P_B}{\sqrt{\frac{P_A(1-P_A)+P_B(1-P_B)}{n}}},$$

alors $U \sim \mathcal{N}(0,1)$. On peut donc construire un test au risque d'erreur α par la région de rejet R_α définie par

$$R_\alpha := \left\{ \mid U \mid > q_{1-\alpha/2} \right\},$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

Dans la pratique on comparera les trois groupes deux à deux. On pourrait également analyser les proportions sur les trois groupes en même temps avec un test du chi-deux mais les résultats seraient moins interprétables. En pratique si on se fixe $n = 200$ et $\alpha = 5\%$ et qu'on observe une proportion d'appel de 30% dans le groupe témoins et de 20% dans le groupe de proactivité personnalisée on obtient une statistique de test de

$$\frac{0,3-0,2}{\sqrt{\frac{0,3(1-0,3)+0,2(1-0,2)}{200}}} = 2,325 > 1,960 = q_{0,975},$$

et on rejette l'hypothèse nulle et on conclut que la proactivité a eu un impact sur le comportement des assurés.

4.2.2 Résultats

Au moment du rendu de ce mémoire les résultats du test n'étaient pas encore disponibles. Le test se déroule sur plusieurs week-ends consécutifs et il est possible d'espérer que les résultats permettront d'apprécier les gains potentiels de mesures de proactivité à échelle plus générale. On apportera également de l'importance à l'évolution des résultats au cours des semaines en espérant ne pas constater une trop forte variabilité. Si la proactivité écrite standardisée ou personnalisée permet de sensiblement diminuer la charge d'appels de gestion cela pourrait susciter des réflexions plus générales sur la visibilité donnée à l'assuré sur la gestion de son sinistre. En effet, on sait qu'une forte probabilité d'appel est principalement induite par une activité récente et la proactivité pourrait donc être réalisée au moment même de cette activité.

Il est à noter qu'en cas de résultats concluants le gain en coût de gestion pourrait être conséquent. On peut voir sur la figure 4.2 que la proactivité écrite semble rester rentable jusqu'à une probabilité d'appel de 0,08 et avec un champ d'action aussi large on peut donc raisonnablement espérer éviter plus de 1500 appels réels par semaine si l'on intègre le fait que lors de notre étude seule une partie des appels de gestion a été détectée et qu'un dossier appelant peut susciter plus d'un appel. En effet, pour éviter 1500 appels réels il faut éviter environ 900 dossiers en intégrant la possibilité d'appeler plusieurs fois et 900 dossiers appelants réellement devrait correspondre à un peu plus de 500 dossiers détectés dans nos données. Ramener au coût moyen d'un appel cela représenterait une économie de plus d'un million d'euros à l'année sur le seul périmètre de cette étude avec une perspective d'application à un portefeuille plus large sur des sinistres des périmètres auto ou professionnel.

4.3 Autres utilisations possibles

Dans cette dernière section on se propose de présenter quelques utilisations supplémentaires des résultats découverts au cours de ce mémoire.

4.3.1 Simulation du processus de gestion

Bien que les résultats de ce mémoire ne seraient pas suffisants pour y parvenir, la conception d'un simulateur de trajectoire de gestion peut avoir de nombreux intérêts. En effet, disposer de simulations

sur l'évolution de la gestion et l'indemnisation d'un dossier devrait permettre de modéliser correctement les coûts de gestion mais surtout la répartition temporelle de la charge de gestion en termes de coûts et de besoin en capacité de traitement.

Associée à une modélisation de la sinistralité qui comporte également une dimension temporelle, cette modélisation de la gestion pourrait permettre de disposer d'un outil qui serait en capacité de simuler de manière assez complète le comportement d'un portefeuille de contrat. On peut en effet imaginer une modélisation qui inclut la survenance des sinistres, le coût et la nature des dégâts, les échanges pour la gestion du sinistre, le choix du mode gestion et donc le coût de gestion. Un tel outil pourrait être utile en tarification ou en provisionnement mais aussi en planification de l'activité.

4.3.2 Pistes d'amélioration du processus d'indemnisation

Si l'on se replace dans une optique de compréhension et de réduction de l'arrivée d'appels de gestion, certains résultats de ce mémoire peuvent mettre en avant des pistes d'amélioration du processus d'indemnisation et de gestion au-delà de la proactivité.

Importance de la déclaration

Que ce soit en termes de statistiques descriptives ou de modélisations on a pu constater que les premiers jours de la gestion d'un dossier étaient nettement plus susceptibles de contenir des appels de gestions. Ce constat laisse penser qu'un travail spécifique sur la déclaration pourrait permettre de de mieux contenir la charge d'appels de gestion. Il est possible de considérer qu'au moins une partie des appels de gestion qui interviennent tôt (notamment les premiers jours) dans la vie d'un dossier sont le fruit d'un manque d'information donnée à l'assuré, d'une information non comprise par l'assuré, d'un manque de visibilité sur le déroulé de la gestion ou encore d'une mauvaise communication quant aux canaux de contact disponibles et à privilégier.

Pour évaluer les gains potentiels à réaliser sur ces axes, plus qu'une modélisation il serait pertinent d'analyser le contenu des échanges ayant lieu au moment de la déclaration et les jours suivant et en particulier les appels de déclaration. Cette analyse peut se faire une écoute humaine mais les progrès réalisés ces dernières années en transcription audio vers du texte peuvent permettre une analyse plus volumineuse si on combine la transcription avec des outils de fouille de textes ou de traitement automatique du langage.

Analyse du discours du gestionnaire

Au-delà du fait que les appels soient plus concentrés après l'ouverture, on a pu constater que les différents échanges téléphoniques (déclaration, gestion, et appels sortants) ne viennent pas réduire l'arrivée future d'appels de gestion et même au contraire. Cette remarque a surpris et interrogé sur ces successions d'échanges téléphoniques. On s'attendait à ce que les échanges téléphoniques en apportant de l'information à l'assuré limitent son besoin de rappeler l'assureur.

Des explications possibles de ces répétitions d'appels peuvent venir des mêmes points que ceux évoqués pour la déclaration. À cela on peut supposer à l'aide de dires d'experts basés sur des observations que certains appels se terminent par une formule de politesse invitant l'assuré à ne pas hésiter à recontacter l'assureur. On peut également penser que si le test évoqué plus tôt dans ce chapitre confirme l'efficacité d'une proactivité écrite il pourrait être pertinent de terminer les échanges téléphoniques par un récapitulatif écrit systématique. Une fois de plus l'utilisation d'outils basés sur les avancées technologiques en matière d'Intelligence Artificielle et particulièrement en IA générative

pourrait se révéler utile. Il est en effet envisageable de fournir un résumé des informations délivrées lors d'un appel avec des informations sur la suite des actions à mener grâce à ces outils.

Choix entre communication écrite et téléphonique

Le fait qu'au moins théoriquement la communication sortante ait été identifiée comme préférable à l'écrit plus qu'au téléphone peut également amener à des évolutions de pratiques dans la gestion. En effet, on peut donc recommander que lorsque le gestionnaire à besoin de contacter l'assuré, s'il ne juge pas indispensable de faire ce contact à l'oral, il privilégie le contact par écrit. Si l'on peut concevoir que dans certains cas, surtout les plus complexes, le contact téléphonique puisse être indispensable, il parait assez simple de rendre systématique certains contacts par écrit comme notamment la confirmation de réception de documents.

Plus généralement, dans une situation où l'assuré a reçu des informations précises sur le déroulé de la gestion de son dossier, sauf élément venant modifier la planification de cette gestion, il est concevable que l'ensemble des échanges se fassent à l'écrit. Cette mesure peut a minima être prise pour les assurés qui préfèrent le canal écrit pour communiquer.

4.4 Conclusion

Ce dernier chapitre permet, même en l'absence des résultats du test de proactivité, d'identifier de nombreuses applications des résultats obtenus au cours de ce mémoire. De façon générale, une meilleure compréhension du phénomène de gestion ne peut être que valorisable que ce soit d'un point de vue de l'efficacité opérationnelle et la maîtrise des coûts mais aussi du point de vue de l'assuré qui est en demande d'une gestion et d'une indemnisation claire et efficace.

Bien qu'en attente de confirmation par les résultats du test, ce dernier chapitre devrait également permettre une évolution assez significative dans le paradigme utilisé en gestion. En effet, jusqu'ici lorsque l'assureur essayait de réduire sa charge d'appels entrants en donnant de la visibilité aux assurés cela se faisait presque systématiquement par des appels sortants. Ce mémoire devrait permettre une remise en question de ces pratiques ce qui peut aller avec l'utilisation d'outils numériques pour faciliter la communication entre assureur et assuré sans avoir besoin de la simultanéité des disponibilités.

Conclusion

L'objectif de l'étude menée au cours de ce mémoire était d'étudier, modéliser, comprendre, prédire et piloter le processus de gestion et d'indemnisation afin d'éclairer l'assureur sur ce phénomène aléatoire qui impacte la maîtrise des coûts de gestion et celle de l'activité ainsi que la satisfaction client. Il avait été identifié que les appels de gestion représentaient un aspect stratégique dans ce processus de gestion et d'indemnisation dans la mesure où ces derniers représentent une charge importante que ce soit financièrement mais également en termes de pénibilité tant pour l'assureur que l'assuré.

La modélisation de l'arrivée d'appels de gestion aura été réalisée de manière assez satisfaisante en utilisant des outils de data science appliqués à la fois sur les caractéristiques du dossier de sinistre mais également sur l'historique des événements le concernant. L'utilisation de différents modèles aura permis de montrer qu'il est possible d'affecter une probabilité d'appel dans un futur proche de manière cohérente en conservant des temps de calcul tout à fait acceptables même avec les modèles les plus complexes. Ces modélisations auront montré que la probabilité d'appel était déterminée principalement par la récence de l'activité, bien plus que les caractéristiques du sinistre.

Ce constat a renforcé l'idée de la nécessité de s'intéresser de plus près au processus de gestion et d'indemnisation dans sa globalité y compris dans l'objectif de renforcer la compréhension de l'arrivée des appels de gestion. S'il était assez clair que l'activité d'un dossier comportait des phases d'auto-excitation il aura été compliqué de capter ce phénomène sur l'intégralité des événements à l'aide de processus de Hawkes. En revanche, cette approche aura permis de mettre en avant une adaptation ciblée sur les appels de gestion par la construction d'une intensité d'appel pour chaque dossier à chaque instant. Cette intensité aura notamment permis de mettre en évidence une capacité théorique de la communication écrite à freiner l'arrivée d'appels. L'approche consistant à étudier le processus de gestion et d'indemnisation dans sa globalité en regroupant certains événements dans des séquences semblait prometteuse mais nécessiterait d'être approfondie pour disposer d'une modélisation plus complète.

Les résultats obtenus au cours de l'étude auront permis de mettre en avant des pistes d'amélioration du processus de gestion et d'indemnisation à commencer par l'identification d'une réduction potentielle de la charge de gestion en utilisant une proactivité à l'écrit. Cette hypothèse devra être vérifiée par une expérimentation en conditions réelles. On aura également pu identifier tout au long du mémoire certains points qui laissent à penser qu'il existe des leviers d'amélioration dans l'activité de gestion et d'indemnisation notamment autour des appels et de la déclaration.

Ce mémoire comprend de nombreuses limites à commencer par l'inexactitude sur la principale variable d'intérêt, les appels de gestion étant à l'heure actuelle détectés de manière assez incomplète. Cette incomplétude ne devrait en revanche pas remettre en cause la nature des conclusions. On peut également citer comme autre limite le fait que le contenu des échanges entre l'assureur et l'assuré n'a

que peu été étudié bien qu'il devrait permettre des modélisations plus performantes mais bien plus compliquées à mettre en œuvre.

Ce mémoire aura permis de mettre en avant les avantages potentiels de la mise en place d'analyses actuarielles sur des activités de la chaîne de valeur des produits d'assurances qui sont habituellement exclues du domaine. En effet, l'assureur a tout intérêt à aborder l'ensemble de son activité sous le prisme du risque y compris dans les domaines afférents aussi au marketing, à la relation client ou au pilotage d'activité et de ne pas limiter l'expertise actuarielle aux domaines dans lesquels elle est usuellement employée.

Bibliographie

Bacry, E., Bompaire, M., Gaïffas, S. et Poulsen, S. (juill. 2017). tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*. eprint : 1707.03003.

Bacry, E., Mastromatteo, I. et Muzy, J.-F. (fév. 2015). Hawkes processes in finance. Papers 1502.04592. arXiv.org. URL : https://ideas.repec.org/p/arx/papers/1502.04592.html.

Breiman, L. (2001). Random Forests. *Machine Learning* 45.1, p. 5-32.

Breiman, L., Friedman, J., Stone, C. J. et Olshen, R. (1984). Classification and Regression Trees. Chapman et Hall/CRC.

Brodeur, N. (2019). Chaîne d'indemnisation IRD. Mémoire d'actuariat. Covéa : Centre d'études actuarielles.

Chen, T. et Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA : ACM, p. 785-794.

HAWKES, A. G. (avr. 1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58.1, p. 83-90. eprint : https://academic.oup.com/biomet/article-pdf/58/1/83/602628/58-1-83.pdf.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781.

Nelder, J. A. et Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)* 135.3, p. 370-384.

Norris, J. R. (1997). Markov Chains. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, p. 2825-2830.