UFR Mathématique Informatique Strasbourg

duas
Diplôme Universitaire des Actuaires de Strasbourg

Mémoire présenté pour l'obtention du DUAS et l'admission à l'Institut des Actuaires

le 16/12/2022

Par : LARDE Quentin

Titre: Remboursements à tort en assurance santé : Application des algorithmes de classification supervisée à la recherche de règlements en double

Confidentialité : ☐ NON ☐ OUI Durée : ☒ 1 an ☐ 2 ans ☐ 3 ans ☐ 4 ans ☐ 5 ans

Membres du jury de l'IA :

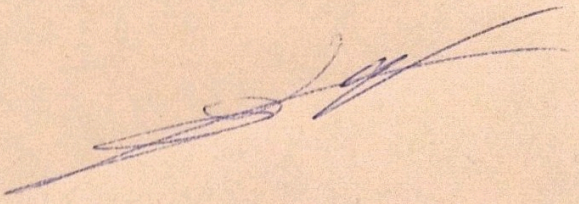
D. DUBOIS
L. FONTAINE

Membres du jury de l'Unistra :

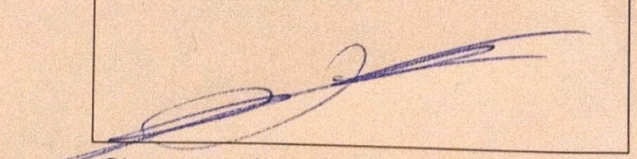
J. BERARD
E. BIRMELE
A. COUSIN
P.-O. GOFFARD
M. MAUMY-BERTRAND

Entreprise :
Assurances du Crédit Mutuel

Directeur de mémoire (entreprise) :
Nom :LAURENT Xavier

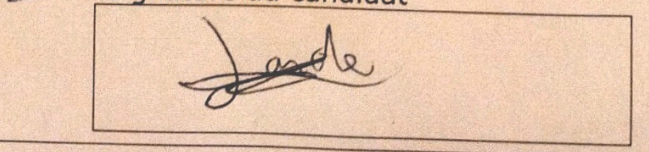


Signature du responsable entreprise



Secrétariat : Mme Stéphanie Richard

Signature du candidat



Remerciements

Je souhaiterais tout d'abord remercier Serge RUDIO qui a pris le temps de répondre à mes questions, de me conseiller et grâce à qui j'ai pu parfaitement m'intégrer au sein du service. Je tiens à remercier Xavier LAURENT qui m'a fait confiance en me permettant de travailler sur des sujets variés et intéressants.

Je tiens à exprimer ma gratitude envers l'équipe du Contrôle Permanent de Strasbourg pour leur disponibilité, leur expérience et leurs précieux conseils qui ont été un atout indéniable permettant de passer de nombreux caps.

Je tiens à remercier l'équipe de gestion santé pour leurs explications et le temps accordé pour m'aider à comprendre les particularités de l'assurance santé et les mécanismes de gestion. Je voudrais également remercier les équipes de lutte anti-fraude qui m'ont aiguillé dans différents travaux et apporté de précieux conseils.

Merci au corps professoral du diplôme d'actuariat qui m'a permis d'acquérir les compétences requises dans le domaine de l'actuariat et de la *data science*.

Enfin, je remercie Alexis REBERGUES et Kim POUILLY pour leur sens critique et leur aide à la relecture du mémoire.

Résumé

Parmi les secteurs de l'assurance, la santé est celui qui correspond au plus grand volume de prestations. Afin de diminuer les coûts et la durée des traitements, la gestion des sinistres santé doit être rapide, ce qui peut conduire à des remboursements indus. Les automatisations nécessaires ont néanmoins des aspects positifs : elles permettent de collecter beaucoup de données.

L'utilisation de la *data* collectée peut permettre de détecter des pertes d'argent liées à des paiements à tort. Pour être optimale, cette détection doit être effectuée *a priori*, autrement dit, avant tout remboursement.

L'objectif du mémoire est de proposer des méthodes de détection des remboursements à tort et en particulier dans le cadre des remboursements effectués en double. Une base de données est construite pour des contrats, bénéficiaires, dates de soins et montant de frais réels égaux au niveau des lignes de soins.

Le but est de prouver que des algorithmes de classification supervisée peuvent être appliqués à ce type de problème. L'utilisation d'indicateurs de performance dans le cadre des algorithmes de classification permet de sélectionner le modèle qui convient le mieux.

L'algorithme sélectionné est appliqué tous les jours pour bloquer les paiements à tort de soins saisis la veille. En l'appliquant sur l'historique, un nombre important d'indus est détecté. Cela permet d'enrichir les solutions de détections a priori actuellement en place et d'estimer le risque lié à ces indus.

Mots-clés : indus, doublons, doubles règlements, santé, *machine learning*, *data science*, apprentissage supervisé

Abstract

Among the insurance sectors, health accounts for the largest volume of benefits. In order to reduce the costs and duration of treatment, the management of health claims must be rapid, which can lead to undue reimbursements. The necessary automations have nevertheless positive aspects : they allow to collect a lot of data.

The use of the collected data can detect money losses related to undue payment. To be optimal, this detection must be carried out a priori, in other words, before any reimbursement.

The objective of the study is to propose methods of detecting undue payments and in particular in the context of duplicate payments. A database is built by contract, beneficiary, date of care and amount of actual costs at the level of the act of care.

The aim is to prove that supervised classification algorithms can be applied to this type of problem. The use of performance indicators in classification algorithms help to select the most suitable model.

The selected algorithm is applied daily to block wrongly entered care payments the day before. By applying it to history, a large number of indus is detected. This makes it possible to enrich the a priori detection solutions currently in place and to estimate the risk associated with these indus.

Keywords : undue payments, duplicate payments, health insurance, machine learning, data science, supervised learning

Note de synthèse

Parmi les secteurs de l'assurance, la santé est celui qui correspond au plus grand volume de prestations. Afin de diminuer les coûts et la durée des traitements, la gestion des sinistres santé doit être rapide, ce qui peut conduire à des remboursements indus. De plus, l'assureur peut obtenir l'information d'un soin de différentes sources. Cette information peut provenir de la sécurité sociale, du tiers payant ou de l'assuré. C'est pourquoi le gestionnaire doit disposer de différentes alertes lui donnant des indications pour éviter les indus. Il peut s'agir d'alertes qui indiquent que le soin a probablement déjà été saisi, d'alertes indiquant des vérifications qui doivent être faites avant une validation du remboursement. L'existant en matière de double règlement consiste en deux étapes : une détection *a priori* avec des alertes à la saisie du soin et une détection *a posteriori* le lendemain des soins qui permet de bloquer le virement en cas d'indu.

Problématique du mémoire

L'objectif du mémoire est de proposer des méthodes de détection des remboursements à tort en particulier dans le cadre des remboursements effectués en double sur un périmètre défini. Cette étude permet d'améliorer la procédure de détection des règlements *a posteriori*. De plus, l'analyse des cas permet de trouver des failles dans la détection *a priori*. Pour cela, des algorithmes de classification supervisée sont utilisés. Une estimation du montant annuel du risque est effectuée à la fin du mémoire.

Méthodologie suivie

Pour construire la base de données, des couples de lignes de soins avec des caractéristiques proches sont analysés. La base de données est constituée de lignes de soins pour lesquelles la date de soins, le bénéficiaire et le coût réel des soins sont identiques et dont le règlement est supérieur à 50€.

Ensuite, il s'agit de définir si les soins du périmètre considéré correspondent à des règlements en double ou non. La base de données est construite à partir de cas pour lesquels une ligne de soins a été régularisée (annulation de règlement, dette pour l'assuré) et d'autres sans régularisation. Différents indicateurs sont ajoutés à la base de données en utilisant les informations obtenues à propos des lignes de soins mises en relation. Les indicateurs créés permettent de retraiter la base et de considérer certains cas comme doublons et d'autres comme non doublons.

Des algorithmes de classification non-supervisée sont ensuite appliqués à la base de données retraitée. Les *Support Vector Machines*, les *Random Forest*, le *Gradient Boosting*, les réseaux de neurones et le *Stacking* sont utilisés. Pour discriminer les modèles, les matrices de confusion et la courbe ROC sont utilisés.

Une fois l'obtention des cas de règlements en double, l'adéquation de lois de probabilité au nombre annuel de règlements en double et au coût unitaire de ces doublons est utilisé pour estimer le risque annuel.

Travaux

Retraitement de la base de données

Après obtention de la base de données avec les étapes décrites, elle est retraitée à l'aide des indicateurs créés. Finalement, elle est composée de 60,5% de lignes en doublon. La plus grande partie des soins correspond à de l'hospitalisation et du dentaire qui sont des postes très soumis au risque de règlement en double du fait de la procédure de gestion et du risque de double télétransmission d'un même soin.

Application des algorithmes de classification supervisée

L'application des algorithmes de classification supervisée donne des résultats proches en termes d'*accuracy* et de précision. Les réseaux de neurones semblent être un bon compromis entre performances et temps de calcul. Les indicateurs suivants sont obtenus :

Base d'apprentissage Quantile 0,5		Prédiction	
		0	1
Valeur	0	8124	483
	1	413	12811

Indicateurs base d'apprentissage	
Accuracy	95,90%
Taux de vrais positifs	96,88%
Taux de vrais négatifs	94,39%
Taux de faux positifs	5,61%
Précision	96,37%

Base de test Quantile 0,5		Prédiction	
		0	1
Valeur	0	2042	110
	1	111	3195

Indicateurs base test	
Accuracy	95,95%
Taux de vrais positifs	96,64%
Taux de vrais négatifs	94,89%
Taux de faux positifs	5,11%
Précision	96,67%

Matrice de confusion et performance du réseau de neurones

Les *Random Forest* et le *Gradient Boosting* fournissent des résultats similaires au réseau de neurones mais sont bien plus coûteux en temps.

Le *stacking* est également considéré. Il consiste à utiliser les prédictions d'autres modèles pour constituer une prédiction plus robuste.

Les trois modèles de la première couche choisis sont un modèle *Random Forest*, Un modèle *XGBOOST* et un réseau de neurones. L'*accuracy* et la précision sont les meilleures de tous les modèles présentés. Cependant, l'utilisation de ce modèle requiert l'apprentissage de tous les autres modèles et est donc bien plus coûteux en temps.

Base d'apprentissage Quantile 0,5		Prédiction 0	Prédiction 1
Valeur	0	8247	360
	1	381	12843

Indicateurs base d'apprentissage	
Accuracy	96,61%
Taux de vrais positifs	97,12%
Taux de vrais négatifs	95,82%
Taux de faux positifs	4,18%
Précision	97,27%

Base de test Quantile 0,5		Prédiction 0	Prédiction 1
Valeur	0	2066	86
	1	121	3185

Indicateurs base test	
Accuracy	96,21%
Taux de vrais positifs	96,34%
Taux de vrais négatifs	96,00%
Taux de faux positifs	4,00%
Précision	97,37%

Matrice de confusion et performance du *stacking*

Finalement, sur le périmètre considéré, le réseau de neurones est conservé.

Estimation du risque annuel sur le périmètre de détection

À l'aide de l'algorithme obtenu avec les réseaux de neurones, une base de données des doubles règlements potentiels est construite pour les décomptes saisis entre 2018 et 2022 sans régularisation.

Le risque annuel est modélisé de la manière suivante :

$$Risque = \sum^{N} DR_k \times B_k$$

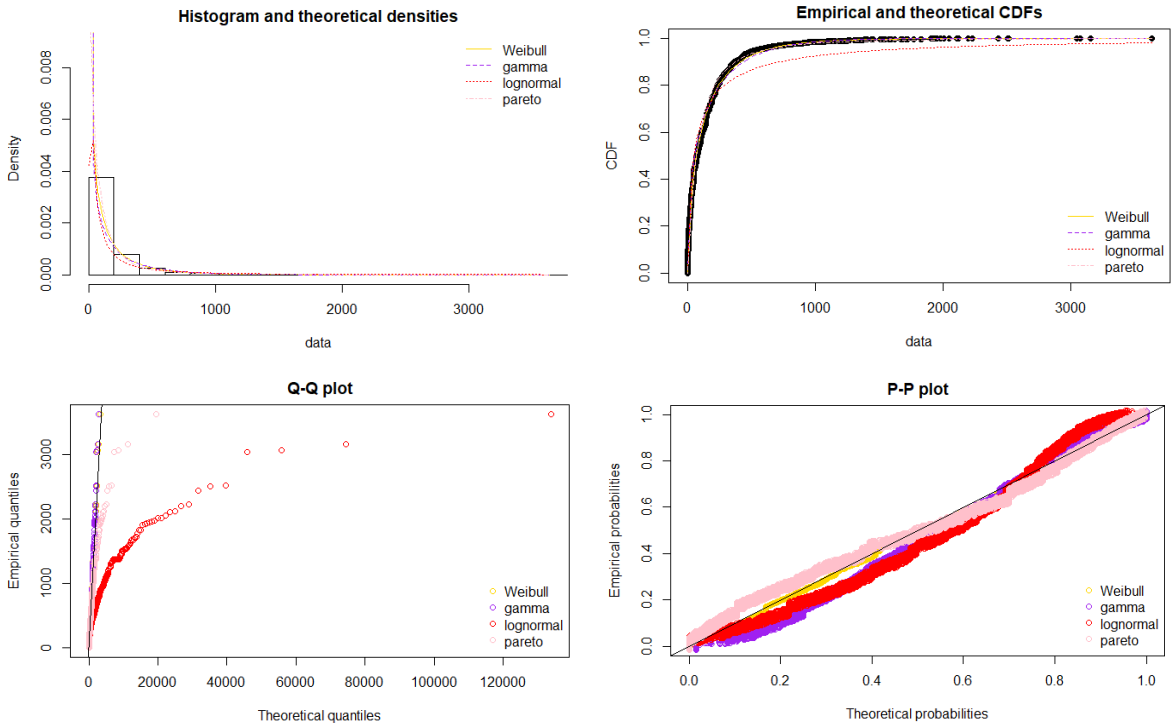
Avec :

- N une variable aléatoire qui modélise le nombre de règlements en double potentiels par an,
- (DR_k) une suite de variables aléatoires indépendantes identiquement distribuées qui modélisent l'impact d'un double règlement potentiel,
- (B_k) une suite de variables aléatoires indépendantes identiquement distribuées modélisant la précision de l'algorithme,
- $N, (DR_k), (B_k)$ indépendantes.

En prenant en compte le nombre décomptes saisis entre 2019 et 2021 de la base de données, en moyenne 1459 décomptes avec des règlements en double potentiels sont saisis par an sans régularisation. La variance de ce nombre est 8662 ce qui conduit à sélectionner une loi binomiale négative pour la variable N.

Pour approcher la loi le (DR_k), l'adéquation de lois de probabilités Weibull, log-normale, Pareto et Gamma à $(DR_k - 49{,}9)$ est analysée à l'aide du package *fitdistrplus* de *R*.

Pour mesurer graphiquement l'adéquation des lois, les graphiques suivant sont utilisés :



Graphiques de comparaison entre les données empiriques et les lois de probabilité sélectionnées

Les statistiques suivantes sont obtenues :

```
Goodness-of-fit statistics
                              weibull       gamma    lognormal      pareto
Kolmogorov-Smirnov statistic  0.03776136   0.1031462   0.09475285  0.06626424
Cramer-von Mises statistic    1.29482377  20.3800444  25.16865323  8.23873415
Anderson-Darling statistic   13.31788533 120.9965267 170.27538350 81.08059793

Goodness-of-fit criteria
                                 weibull    gamma lognormal   pareto
Akaike's Information Criterion 78616.76 79006.32  80848.92 79409.23
Bayesian Information Criterion 78630.37 79019.93  80862.52 79422.83
```

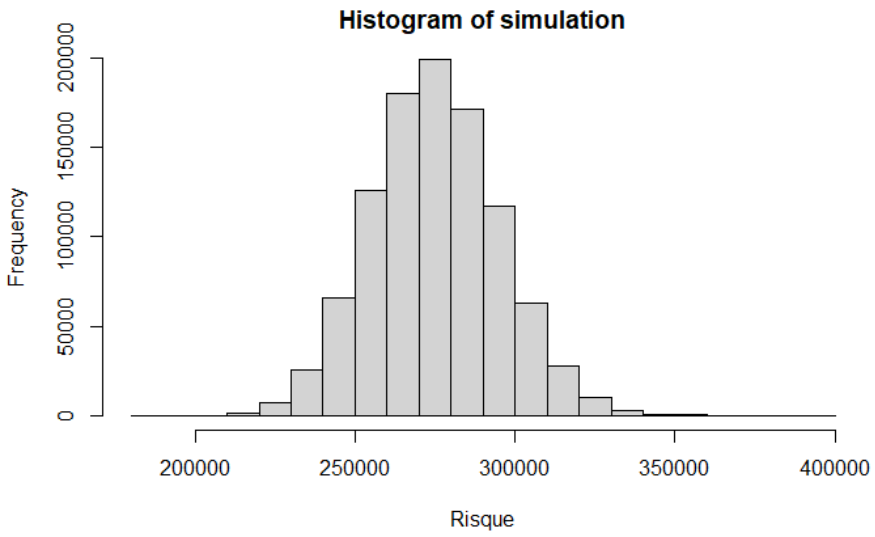
Valeurs des statistiques permettant de comparer l'adéquation des lois de probabilité sélectionnées aux données

La loi de Weibull est alors choisie.

Pour modéliser le fait qu'un décompte de soins considéré comme doublon par l'algorithme peut ne pas être forcément lié à un indu, une loi de Bernoulli est utilisée. Le paramètre de la loi sera alors 90% étant donné la phase de test de l'algorithme par les gestionnaires qui conduit vers une précision de 90%.

Il est désormais possible de simuler la valeur du risque.

Avec 1 000 000 de simulations, l'histogramme du risque est tracé.



Histogramme des simulations du risque

Ainsi, $E(Risque) = 275\,000$ *et* $VaR_{99,5\%}(Risque) = 330\,000$€.

Limites de l'étude

Le retraitement de la base de données ne peut pas être parfait. En effet, parmi les cas avec régularisation, il est impossible de savoir avec une précision parfaite quels cas concernent des doubles règlements. De plus, parmi les cas sans régularisation, il est impossible de contrôler tous ces cas.

L'estimation du risque est limitée par les évolutions du portefeuille et des méthodes de détections au fil des années qui peuvent faire varier les indus. C'est pour cela qu'un nombre d'années limité a été choisi.

Conclusion

Le volume de prestations et les différentes manières d'obtenir l'information d'un soin en assurance santé résulte naturellement en des doublons de règlement. Le but de ce mémoire est d'utiliser la classification supervisée pour détecter des doubles règlements.

Pour cela une base de données avec des lignes de soins proches a été construite et un indicateur permettant de classifier les lignes en doublon ou non a été obtenu à l'aide de différents indicateurs.

Les algorithmes de classification supervisée ont ensuite été appliqués sur cette base. Les résultats de l'application des algorithmes semblent faire apparaître une problématique de sur-apprentissage pouvant s'expliquer par la construction de la base de données. En effet, le fait qu'il n'est pas possible de savoir sans contrôle si une ligne fait référence à un règlement en double ou non a conduit à utiliser des règles métier pour créer l'indicateur doublon. Cela conduit à un biais sur le modèle qui est forcément parfait par construction. La base de données pourra être mise à jour régulièrement avec les indus réellement topés, en supprimant petit à petit les règles métiers une fois la volumétrie suffisante.

Cette étude a permis de récupérer des indus importants sur ce périmètre de règlements en double. Elle permet chaque jour de détecter des cas de règlements en double saisis la veille pour bloquer des paiements indus.

Sur le périmètre choisi, l'adéquation de lois de probabilités au nombre d'indus potentiels et au coût unitaire permet d'estimer le montant du risque et d'en donner une *Value At Risk* à 200 ans.

Executive summary

Among the insurance sectors, health accounts for the largest volume of benefits. In order to reduce the costs and duration of treatment, the management of health claims must be rapid, which can lead to undue reimbursements. In addition, the insurer can obtain information from various sources. This information may come from social security, the third-party paymend or the insured person. That is why the manager must have various alerts giving him indications to avoid undue hardship. These can be alerts that indicate that care has likely already been entered, alerts that indicate checks that must be done before a refund validation. The existing dual settlement system consists of two stages: a priori detection with alerts at the entry of care and a posteriori detection of care the day after which the transfer can be blocked in case of undue payment.

Purpose of the study

The purpose of the study is to propose methods for detecting undue payments, particularly in the context of duplicate payments over a defined scope. This study makes it possible to improve the procedure for detecting regulations a posteriori. In addition, case analysis makes it possible to find flaws in a priori detection. For this, non-supervised classification algorithms are used. An estimate of the annual amount of risk is made at the end of the study.

Methodology

To build the database, couples of care lines with similar characteristics are analyzed. The database is made up of lines of care for which the date of care, the beneficiary, the actual amount of care are identical and whose payment is more than 50€.

Then, it is a matter of defining whether the care of the perimeter in question corresponds to duplicate regulations or not. The database is built from cases for which a line of care has been regularized (cancellation of payment, debt for the insured) and others without regularization. Different indicators are added to the database using the information obtained about the connected care lines. The indicators created make it possible to reprocess the base and to consider some cases as duplicate payments and others as not duplicate payments.

Supervised classification algorithms are then applied to the reprocessed database. Support Vector Machines, Random Forest, Gradient Boosting, neural networks and Stacking are used. To discriminate between models, confusion matrices and the ROC curve are used.

Once duplicate payments cases are obtained, the adequacy of probability laws to the annual number of duplicate payments and the unit cost of those duplicates is used to estimate the annual risk.

Approach

Reprocessing of the database

After obtaining the database with the steps described, it is reprocessed using the indicators created. Finally, it is composed of 60,5% duplicate lines. Most of the care corresponds to hospitalization and dentistry, which are positions which are highly exposed to the risk of double payments due to the management procedure and the risk of double teletransmission of the same care.

Application of supervised classification algorithms

The application of supervised classification algorithms gives close results in terms of accuracy and precision. Neural networks seem to be a good compromise between performance and computation time. The following indicators are obtained:

Training set Quantile 0,5		Prediction	
		0	1
Value	0	8124	483
	1	413	12811

Training set indicators	
Accuracy	95,90%
True positive rate	96,88%
True negative rate	94,39%
False positive rate	5,61%
Precision	96,37%

Test set Quantile 0,5		Prediction	
		0	1
Value	0	2042	110
	1	111	3195

Test set indicators	
Accuracy	95,95%
True positive rate	96,64%
True negative rate	94,89%
False positive rate	5,11%
Precision	96,67%

Confusion matrix and neural network performance

Random Forest and Gradient Boosting provide similar results to the neural network but are much more time-consuming.

Stacking is also considered. It involves using predictions from other models to provide a more robust prediction.

The three first layer models chosen are a Random Forest model, an XGBOOST model and a neural network. Accuracy and precision are the best of all models presented. However, using this model requires learning all other models and is therefore much more time-consuming.

Training set Quantile 0,5		Prediction	
		0	1
Value	0	8247	360
	1	381	12843

Training set indicators	
Accuracy	96,61%
True positive rate	97,12%
True negative rate	95,82%
False positive rate	4,18%
Precision	97,27%

Test set Quantile 0,5		Prediction	
		0	1
Valeur	0	2066	86
	1	121	3185

Test set indicators	
Accuracy	96,21%
True positive rate	96,34%
True negative rate	96,00%
False positive rate	4,00%
Precision	97,37%

Confusion matrix and stacking performance

Finally, on the perimeter considered, the neural network is selected.

Annual risk estimate for the detection scope

Using the algorithm obtained with neural networks, a database of potential double payments is built for the counts entered between 2018 and 2022 without regulation.

The annual risk is modelled as follows :

$$Risk = \sum^{N} DP_k \times B_k$$

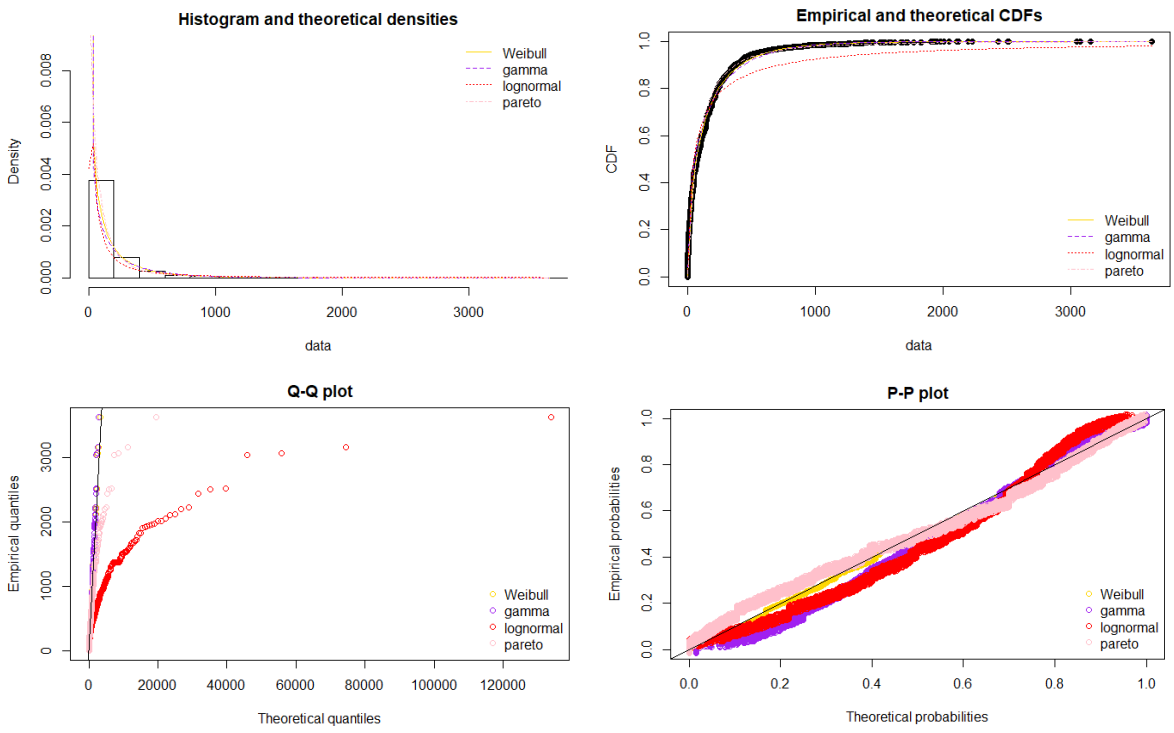
With :

- N a random variable that models the number of potential duplicate payments per year,
- (DP_k) a suite of independent identically distributed random variables that model the impact of a potential double payment,
- (B_k) a suite of independent identically distributed random variables modeling algorithm accuracy,
- $N, (DR_k), (B_k)$ independent.

Taking into account the number of counts entered between 2019 and 2021 of the database, an average of 1,459 counts with potential duplicate payments are entered each year without regularisation. The variance of this number is 8662 which leads to select a negative binomial law for the variable N.

To approach the law of (DP_k), he adequacy of the Weibull, log-normal, Pareto and Gamma laws to $(DP_k - 49{,}9)$ is analyzed using the fitdistrplus R package.

To graphically measure the adequacy of laws, the following graphs are used:



Comparison charts between empirical data and selected probability laws

The following statistics are obtained:

```
Goodness-of-fit statistics
                                 weibull       gamma    lognormal      pareto
Kolmogorov-Smirnov statistic  0.03776136   0.1031462   0.09475285  0.06626424
Cramer-von Mises statistic    1.29482377  20.3800444  25.16865323  8.23873415
Anderson-Darling statistic   13.31788533 120.9965267 170.27538350 81.08059793

Goodness-of-fit criteria
                               weibull    gamma lognormal   pareto
Akaike's Information Criterion 78616.76 79006.32  80848.92 79409.23
Bayesian Information Criterion 78630.37 79019.93  80862.52 79422.83
```

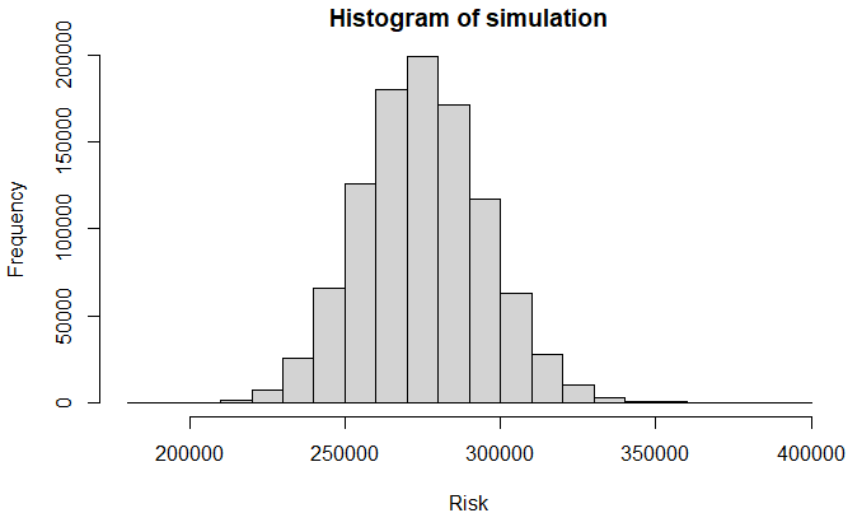
Values of statistics to compare the suitability of the goodness-of-fit to the data

Weibull's law was then chosen.

To model the fact that a count of care considered as a duplicate by the algorithm may not necessarily be linked to an indu, a Bernoulli law is used. The law parameter will then be 90% given the testing phase of the algorithm which leads to an accuracy of 90%.

It is now possible to simulate the value of the risk.

With 1,000,000 simulations, the risk histogram is plotted.



Histogram of risk simulations

Thus, $E(Risque) = 275\,000$ and $VaR_{99,5\%}(Risque) = 330\,000€$.

Study limitations

Reprocessing the database can't be perfect. Indeed, among the cases with regularisation, it is impossible to know with perfect precision which cases concern double payments. Moreover, of the cases without regularization, it is impossible to control all these cases.

The estimation of risk is limited by the evolution of the portfolio and the detection methods over the years which can cause indus. That is why a limited number of years was chosen.

Conclusion

The volume of benefits and the different ways of obtaining information about a health insurance treatment naturally results in duplicate payments. The purpose of this study is to use supervised classification to detect duplicate payments.

For this purpose a database with close care lines was built and an indicator to classify lines in duplicate or not was obtained using different indicators.

The supervised classification algorithms were then applied on this basis. The results of the application of algorithms seem to reveal a problem of overfitting that can be explained by the construction of the database. Indeed, the fact that it is not possible to know without checking whether a line refers to a duplicate payment or not has led to the use of rules to create the duplicate indicator. This leads to a bias on the model which is necessarily perfect by construction. The database can be updated regularly with the detected undue payments, by gradually removing the rules once the volume is enough.

This study has made it possible to recover significant losses in this area of duplicate payments. It allows each day to detect cases of duplicate payments entered the day before to block undue payments.

On the chosen scope, the adequacy of laws of probability to the number of potential indus and to the unit cost allows to estimate the amount of the risk and to give a 200-year Value At Risk.

Table des matières

Introduction

Les indus en assurance résultent en des pertes qu'il est difficile de quantifier car ils ne peuvent pas tous être détectés. Néanmoins, la fraude seule constituerait environ 10% des prestations payées[1] ce qui fait des indus une préoccupation importante pour l'assureur.

En assurance santé, la gestion est de plus en plus automatisée voire industrialisée du fait du grand volume de prestations. De plus, l'assureur peut obtenir l'information d'un soin de différentes sources. Cette information peut provenir de la sécurité sociale, du tiers payant ou de l'assuré. C'est pourquoi le gestionnaire doit disposer de différentes alertes lui donnant des indications pour éviter les indus. Il peut s'agir d'alertes qui indiquent que le soin a probablement déjà été saisi, d'alertes indiquant des vérifications qui doivent être faites avant une validation du remboursement.

L'approche classique pour la détection d'indus est à base de filtres, construits à l'aide de règles déterminées par des experts. Grâce au développement de la *Data Science* et des progrès informatiques qui améliorent le temps de calcul, les nombreuses données en assurance santé peuvent être analysés de manière plus performante. Les méthodes d'apprentissage supervisé et non supervisé peuvent ainsi permettre de calibrer des alertes prenant en compte plus de critères. L'apprentissage supervisé requiert des données pour lesquelles la qualification en tant qu'indu ou non est déjà disponible.

Une base de données est construite sur un historique pour prouver que l'apprentissage supervisé est une solution performante pour détecter les indus. L'utilisation de critères et l'analyse de la base de données permettent de déterminer si les lignes correspondent à des doubles règlements ou non. Les algorithmes d'apprentissage supervisé sont ensuite appliqués.

Le premier chapitre permet de décrire les remboursements à tort en assurance, leur forme, leurs conséquences, et les outils permettant de les éviter. Le chapitre suivant présente l'assurance santé et les nombreuses catégories de soins qui la compose ainsi que son fonctionnement en lien avec le cadre légal en France. Ensuite, les

[1] Source : **https://tribune-assurance.optionfinance.fr/lessentiel/le-potentiel-de-fraude-en-europe-est-estime-a-10-des-prestations-payees.html**

remboursements à tort en assurance santé peuvent être présentés à l'aide de différents exemples. Le dernier chapitre repose sur l'application de l'apprentissage supervisé aux doublons de règlements sur un périmètre défini.

1.Les remboursements à tort en assurance

Il convient tout d'abord de présenter les remboursements à tort. Pour cela, les notions suivantes seront présentées :

- La fraude à l'assurance, les différents types de fraudes et profils de fraudeurs,
- Les remboursements à tort et leurs conséquences sur l'assureur,
- La prescription en assurance qui est essentielle dans le cadre de la récupération des indus,
- Les outils de prévention des indus.

1.1 Présentation de la fraude à l'assurance

1.1.1 Définition de la fraude à l'assurance

La fraude est un acte ou une omission volontaire ayant pour objet de tromper l'assureur. Il en résulte un préjudice moral et ou financier pour l'assureur, le bénéficiaire souhaitant obtenir une économie (minoration de la prime d'assurance), le versement d'une indemnité indue.

La fraude se distingue aisément de la simple erreur d'appréciation ou de l'utilisation assidue d'une garantie assimilable à une sur consommation (exemple : renouvellement annuel du forfait optique).

1.1.2 Les différents types de fraudes

La fraude peut intervenir à toutes les étapes du contrat et peut prendre différentes formes :

- La fraude à la souscription
    1) Sur l'identité du souscripteur ou sur le bénéficiaire de la garantie,
    2) Faux justificatifs (scolarité, valeur assurée …),
    3) Dissimulations de faits (absence d'aléa, sinistre déjà survenu …),
    4) Dissimulations de risques aggravants (invalidité, affection longue durée …).
- La fraude lors d'un sinistre
    1) Faux sinistre : sinistre fictif, monté de toute pièces par l'assuré, le sinistre n'est jamais survenu (faux décès, faux incendie …),
    2) Sinistre ajusté : sinistre réel mais ses conditions et son ampleur sont ajustés à l'existence de garanties précises permettant la prise en charge (tromperie sur les faits, les circonstances, les pièces produites …),

3) Sinistre exagéré : sinistre réel, mais son impact est ajusté à l'existence de garanties et prévient le fait que l'assureur va indemniser sur des bases minorées,
4) Sinistre volontaire : sinistre réel, mais dont la commission est volontaire au sens ou sa réalisation est contrôlée pour obtenir un effet maximum permettant de mettre en œuvre les garanties souscrites (automutilation, entretien volontaire de lésions),
5) Sinistre contenant plusieurs de ces modes opératoires.

- L'excès d'assurance
  1) Sur-assurance,
  2) Assurances multiples.

Un point commun existe entre les fraudes lors d'un sinistre : la production de faux documents ou de faux justificatifs.

Figure 1 – Les différents types de fraude

1.1.3 Les différents profils de fraudeurs

Les fraudeurs peuvent être répartis en trois profils.

Le fraudeur opportuniste va profiter de la survenance d'un sinistre réel pour obtenir une indemnité indue ou surévaluée en modifiant volontairement certains aspects. Il peut ajuster les circonstances, augmenter le préjudice, falsifier des justificatifs.

L'amateur va volontairement adapter les faits d'un sinistre pour s'assurer une prise en charge. Il a souvent des difficultés financières et son sinistre est peu, mal ou pas garanti. Il peut modifier les circonstances, décaler la date, impliquer un tiers complice.

Le « professionnel » ne souscrira un contrat que dans le seul but de percevoir une indemnisation indue. Il sait parfaitement comment frauder. Il est parfois membre d'un réseau, d'une communauté. Il est souvent récidiviste.

Le schéma suivant résume les différents profils.

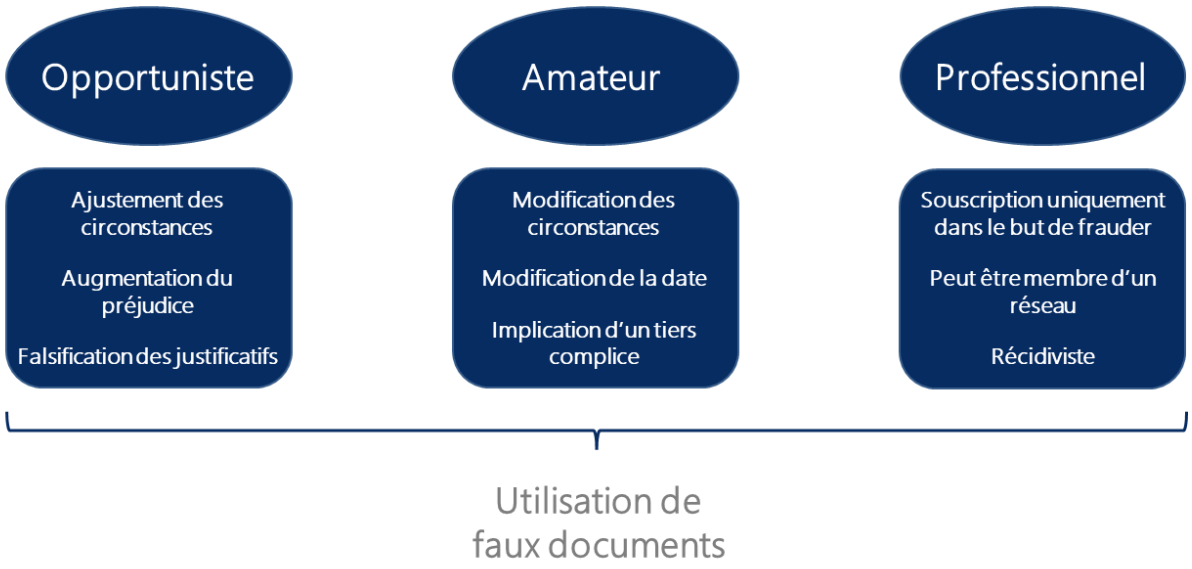


Figure 2 – Les profils de fraudeurs

1.1.4 Les obligations de l'assureur en cas de fraude

D'après l'article L113-8 du Code des Assurances, si l'assuré de mauvaise foi a voulu tromper l'assureur, alors le contrat est nul. Il y a alors disparition rétroactive de la garantie due par l'assureur et le contrat n'est pas opposable aux bénéficiaires du contrat et aux victimes qui agissent par l'action directe en assurance de responsabilité. En guise de dédommagement, l'assureur a le droit de conserver les primes encaissées.

Cependant, la bonne foi de l'assuré est supposée (Article 2 714 du Code civil) donc la mauvaise foi doit être établie. Ainsi, les sanctions de l'article L113-8 du Code des Assurances ne sont applicables que si l'assureur peut prouver la mauvaise foi de l'assuré.

D'après l'article L113-9 du Code des Assurances, la déclaration inexacte de la part de l'assuré dont la mauvaise foi n'est pas établie, n'entraine pas la nullité du contrat d'assurance.

Si l'irrégularité est constatée avant tout sinistre, alors l'assureur a le droit :

- Soit de maintenir le contrat moyennant une augmentation de prime accepté par l'assuré,
- Soit de résilier le contrat 10 jours après notification adressée à l'assuré par lettre recommandée avec accusé de réception en restituant à l'assuré la portion de primes associée à la durée pendant laquelle le risque n'a pas couru.

Si l'irrégularité est constatée après un sinistre, alors on applique la « règle de proportionnalité de taux de prime » :

$$\textbf{indemnité réduite} = \textbf{montant du dommage} \times \frac{\textbf{le taux de prime payé}}{\textbf{taux de prime dû}}.$$

L'assureur peut également résilier le contrat d'assurance.

1.2 Description des remboursements indus

En assurance, les indus peuvent être de deux types. L'assuré peut verser une cotisation qu'il n'aurait pas dû verser, ou l'assureur peut verser à tort une indemnité à l'assuré.

Dans le cadre de ce mémoire, les indemnités versées à tort à un assuré sont étudiées. Ces indus vont impacter le montant des sinistres versés par l'assureur.

En notant :

- S le montant des sinistres versés par l'assureur,
- S^d le montant des sinistres dus,
- S^i le montant des sinistres indument versés,
- N le nombre total de sinistres survenus au sein du portefeuille,
- $X_1, \dots, X_N$ la charge unitaire des sinistres versés, $X_1^d, \dots, X_N^d$ leur part dû et $X_1^i, \dots, X_N^d$ leur part indue.

Alors,

$$S = S^d + S^i = \sum_{k=1}^{N} X_k = \sum_{k=1}^{N} (X_k^d + X_k^i)$$

Avec

$$X_k^d + X_k^i = X_k, 0 \leq X_k^d \leq X_k, 0 \leq X_k^i \leq X_k.$$

Les indus en assurance peuvent intervenir pour différentes raisons :

- En cas de fraude non détectée de la part d'un assuré ou d'un professionnel,
- En cas d'erreur de gestion (doublon de règlement, sinistre non garanti, garantie mal appliquée …),
- En cas de recours.

Soit k le sinistre considéré.

En cas de fraude, il y a déchéance de garanties : $X_k^i = X_k$ et $0 = X_k^d$.

En cas de doublon de règlement, $X_k^i = X_k$ et $0 = X_k^d$ pour le deuxième règlement.

Les recours peuvent être de différents types.

Lorsqu'un assuré est hospitalisé suite à un accident de voiture non responsable, le tiers responsable de l'accident (ou son assureur) devrait prendre en charge les frais. Cela peut alors conduire à un recours. Dans ce cas $X_k^i = X_k$ et $0 = X_k^d$.

En assurance santé, lorsqu'un assuré passe en ALD (affection longue durée), certains soins peuvent être remboursés par l'assurance maladie avec un taux de prise en charge plus important. L'ALD peut être mise en place de manière rétroactive et ainsi un trop versé peut avoir été fait pour différents sinistres ($0 \leq X_k^d$ et $0 < X_k^i$).

1.3 Les conséquences des remboursements à tort sur une compagnie d'assurance

Les remboursements à tort impactent l'assureur. En effet, ces indus augmentent le montant total des sinistres. À ratio sinistre sur prime et montant des sinistres dus égaux, plus les indus sont élevés, plus la prime augmente.

En notant :

- S^d le montant des sinistres dus,
- S^{i_1}, S^{i_2} les montants des sinistres indument versés des assureurs 1 et 2,
- P_1, P_2 les primes des assureurs 1 et 2.

Si les assureurs 1 et 2 ont le même ratio $\frac{S}{P}$, alors $\frac{S^d+S^{i_1}}{P_1} = \frac{S^d+S^{i_2}}{P_2}$ et donc $\frac{S^d+S^{i_1}}{S^d+S^{i_2}} = \frac{P_1}{P_2}$.

Si $S^{i_1} < S^{i_2}$, alors $P_1 < P_2$.

À ratio sinistre sur prime et sinistres dus égaux, l'assureur 1 est alors plus concurrentiel que l'assureur 2.

Les indus influent également sur les risques opérationnels. La fraude externe, c'est-à-dire la fraude par un assuré ou un réseau externe à l'entreprise, fait partie des risques opérationnels.

De plus, les remboursements à tort peuvent affecter l'image de l'entreprise. Du côté des remboursements à tort hors fraude, des sommes vont être demandées au bénéficiaire après indemnisation ce qui va résulter en un mécontentement.

Il peut être difficile de récupérer les indus, surtout pour des montants élevés ou pour une fraude. Dans ce cas, les montants non récupérés ou les intérêts sur les indus pendant la période de recouvrement constituent une perte pour l'entreprise.

1.4 La prescription en assurance

D'après l'article 2219 du Code Civil, la prescription est « un moyen d'acquérir ou de se libérer par un certain laps de temps, et sous certaines conditions déterminées par la loi ». Elle permet à l'assureur ou à l'assuré de se libérer de ses obligations après écoulement d'un délai défini par la loi.

D'après l'article L114-1 du Code des assurances, « Toutes actions dérivant d'un contrat d'assurance sont prescrites par deux ans à compter de l'évènement qui y donne naissance », c'est la prescription biennale.

« Toutefois, ce délai ne court que :

1° En cas de réticence, omission, déclaration fausse ou inexacte sur le risque couru, que du jour où l'assureur en a eu connaissance ;

2° En cas de sinistre, que du jour où les intéressés en ont eu connaissance, s'ils prouvent qu'ils l'ont ignoré jusque-là. »

C'est une disposition d'ordre public à laquelle il ne peut être dérogé : le contrat d'assurance ne peut en modifier la durée. Le principe de prescription biennale s'applique à toutes les actions dérivant du contrat d'assurance. Elle concerne principalement les actions qui opposent l'assuré et l'assureur.

Le point de départ du délai est le lendemain à 0 heure de l'évènement qui fonde l'action.

La prescription peut être interrompue par :

- Une citation en justice,
- La désignation d'un expert,
- L'envoi d'une lettre recommandée avec accusé de réception.

Dès l'évènement interruptif de la prescription, c'est le délai de 2 ans qui recommence à courir, le délai déjà écoulé est supprimé et un nouveau délai de prescription démarre.

En cas d'action pour fausse déclaration à la souscription, le délai court à compter du jour où l'assureur a eu connaissance du caractère mensonger de la déclaration.

Pour les actions en répétition de l'indu : sous réserve que le paiement indu résulte du contrat, le jour où l'assureur a eu connaissance du motif de déchéance est le point de départ du délai de prescription.

Si l'assureur a commis une erreur en versant l'indemnité en raison d'une mauvaise appréciation du contrat, la date de paiement est la date de départ de la prescription.

1.5 Les remboursements à tort d'un point de vue juridique

D'après l'article 1235 du Code Civil, « Tout paiement suppose une dette : ce qui a été payé sans être dû, est sujet à répétition ». Une indemnité d'assurance versée à tort peut donc résulter en une action en restitution.

En effet, d'après l'article 1376 du Code Civil, « Celui qui reçoit par erreur ou sciemment ce qui ne lui est pas dû s'oblige à le restituer à celui de qui il l'a indûment reçu ».

Pour récupérer les sommes versées à tort à la suite d'une fraude, si l'assuré qui a fraudé n'accepte pas d'accord amiable, le seul recours de l'assureur est la voie juridique. L'assureur peut alors effectuer une procédure civile qui relève du droit commun et de la restitution de l'indu ou une action pénale pouvant aboutir à la condamnation de l'assuré qui a fraudé. Dans ce cas, l'assureur pourra récupérer les indemnités versées à tort qui seront alors les préjudices résultant de la fraude. Cependant, pour pouvoir utiliser ces voies, l'assureur doit avoir prouvé la fraude. Sinon, il doit déposer une plainte afin qu'une enquête pénale soit ouverte.

Le détail de prescription est alors compris entre 2 et 6 ans en fonction des types d'actions. Les indemnités versées en cas de fraude peuvent être remboursées au civil. La condamnation du fraudeur à verser des dommages-intérêts pour le préjudice peut être obtenue au pénal.

En cas de procédure pénale, l'assuré fraudeur peut être condamné selon les cas suivants :

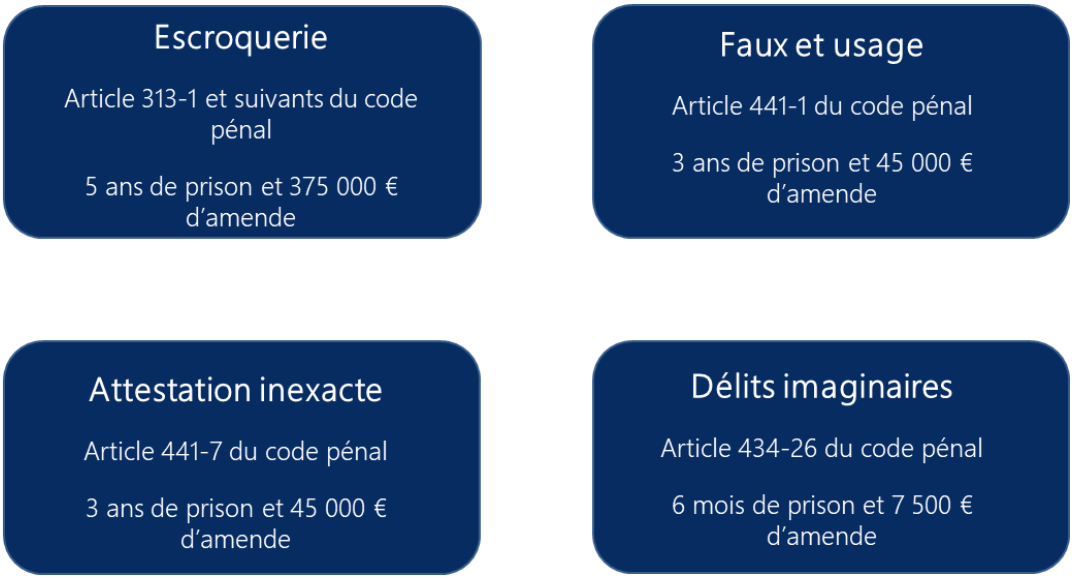


Figure 3 – Les conséquences juridiques de la fraude

Le tableau suivant résume les possibilités de l'assureur en cas de procédure civile ou pénale.

TABLEAU DE SYNTHÈSE DES OPÉRATIONS PROCÉDURALES

	ACTION DEVANT LE JUGE CIVIL	ACTION DEVANT LE JUGE PÉNAL
Mise en œuvre de l'action	• Assignation devant le juge compétent.	• Citation directe devant le tribunal correctionnel. • Plainte devant le Procureur de la République. • Si aucune suite dans le délai de trois mois après le dépôt de plainte simple : plainte avec constitution de partie civile devant le doyen des juges d'instruction.
Délai de prescription	• 5 ans pour l'action en restitution de l'indu (le point de départ devrait être reporté à la date de la découverte de l'existence de l'indu par l'assureur). • 2 ans pour l'action en nullité pour fausse déclaration intentionnelle (le point de départ est reporté au jour où l'assureur a connaissance de la fausse déclaration selon l'article L. 114-1 du code des assurances). • La nullité opposée par voie d'exception est perpétuelle (pas de prescription).	• 6 ans. • À ce jour, le point de départ du délai de prescription est fixé à la date de la commission de l'infraction (évolution possible grâce au nouvel article 9-1 du Code de procédure pénale ?).
Éléments de preuve à réunir	• Tous les éléments nécessaires pour convaincre le juge de l'existence de la fraude (à défaut : les demandes seront rejetées ; condamnation de l'assureur à des dommages-intérêts possible). • Pour l'action en restitution de l'indu en cas d'exagération frauduleuse des pertes : existence et opposabilité d'une clause de déchéance (à défaut, seules les indemnités portant sur l'exagération pourront être restituées). • Pour l'action en nullité : preuve de la mauvaise foi de l'assuré dans la fausse déclaration du risque.	• Pour le dépôt de la plainte : les éléments suffisants pour l'ouverture d'une enquête (en cas de classement sans suite ou de non-lieu, attention au risque de plainte pour dénonciation calomnieuse contre l'assureur). • Devant le tribunal correctionnel : tous les éléments constitutifs de l'infraction.
Demandes possibles	• Restitution de l'intégralité des sommes versées indûment en cas de faux sinistre ou de fausse déclaration à la souscription, ou en cas d'exagération frauduleuse en présence d'une clause de déchéance. • Le cas échéant : nullité du contrat d'assurance pour fausse déclaration intentionnelle.	• Obtention de dommages-intérêts correspondant uniquement au préjudice subi du fait de l'infraction. • Dommages-intérêts divers : préjudice de désorganisation des services, atteinte à l'image, prise en charge des frais d'avocat.

Figure 4 – Possibilités de l'assureur en cas de procédure civile ou pénale[2]

1.6 Les outils de lutte contre les remboursements à tort

Le rôle du gestionnaire est essentiel pour lutter contre les remboursements à tort. En effet, il se trouve en première ligne en étant en contact direct avec les assurés et les documents justificatifs de sinistre. Il va pouvoir détecter des anomalies que ce soit au niveau de doubles règlements, de fraudes, de facturations incohérentes. Cependant, les gestionnaires traitent des volumes de prestations importants et n'ont pas le temps nécessaire à la détection d'anomalies. C'est pourquoi il faut aider les gestionnaires avec différents outils, différentes alertes. Un service de contrôle et de lutte contre la fraude vient ainsi en appui et permettent la mise en place de différentes alertes et d'aider aux traitements avec les différents indus détectés. Différentes alertes pourront alors être

[2] Source : https://www.argusdelassurance.com/acteurs/qu-advient-il-des-indemnites-versees-a-l-assure-a-la-suite-d-une-fraude.118535

mises en place pour détecter des indus résultant de problèmes à la gestion ou de fraudes.

Des outils d'analyse documentaire automatisés sont utiles pour prévenir les remboursements à tort. Ils peuvent permettre de détecter des éléments qui indiquant qu'il y a des incohérences dans les documents. Par exemple, la cohérence d'une carte d'identité, d'un IBAN, des dates sur les documents peut être vérifiée. Une altération ou des modifications effectuées sur un document peuvent être détectées.

L'Agence pour la Lutte contre la Fraude à l'Assurance (ALFA), créée en 1989 a pour but de participer à la lutte contre la fraude à l'assurance. L'ALFA développe des actions de prévention et des moyens de détecter les fraudes. Elle fournit aux entreprises d'assurance des études, de la documentation technique, des listes de fraudeurs détectées afin d'aider les entreprises d'assurance à l'exploitation de scénarios de fraude. Ces fraudes peuvent être liées à des fraudeurs opportunistes, amateurs ou alors des fraudeurs en réseaux.

L'analyse de données et la *data science* sont essentielles dans la prévention de versements indus. Pour cela, les atypies peuvent être analysées pour détecter des erreurs dans les règlements qui peuvent être dus à des problèmes lors de la gestion ou des comportements anormaux de la part des assurés pouvant être de la fraude. Avec l'expérience nécessaire et différents cas détectés, des solutions pourront être mises en place à l'aide d'algorithmes d'apprentissage de type régression ou *machine learning*. Du côté de la gestion, il peut s'agir d'un outil de lutte contre les doubles règlements, du côté de la fraude, il peut s'agir de calibrages pour détecter des fraudeurs selon différents scénarios. L'idéal est d'avoir une détection avant le règlement afin d'éviter toute perte d'argent.

Le volume de prestations important en assurance santé fournit beaucoup de possibilités pour détecter des indus.

2. L'assurance santé

Cette partie a pour but de présenter l'assurance santé et les différents éléments nécessaires à la compréhension de la suite du mémoire. Seront introduits :

- Les actes médicaux,
- Les différents types de soins santé,
- La sécurité sociale et ses composantes,
- Le fonctionnement du régime général dans sa prise en charge des soins,
- Ce qu'est une complémentaire santé et comment fonctionne sa prise en charge,
- Les spécificités santé dans la gestion des documents,
- Une spécificité du Crédit Mutuel : la carte avance santé.

2.1 Les actes médicaux

D'après la définition de la Commission spécialisée de terminologie et de néologie compétente pour le domaine de la santé et le domaine social (2001), un acte médical est un « acte dont la réalisation par des moyens verbaux, écrits, physiques ou instrumentaux est effectué par un membre d'une profession médicale dans le cadre de son exercice et les limites de sa compétence ». Ces actes sont regroupés dans la Classification Commune des Actes Médicaux (CCAM) sous différents codes actes. L'assureur utilise sa propre codification pour les actes de soins qui a pour but d'être adaptée à la saisie des prestations au moment de la gestion.

Les actes de soins sont saisis sous une ligne de soins. Les différentes lignes de soins liées à une visite chez un professionnel de santé sont regroupées sous un décompte de soins. L'assureur effectue ses règlements en se basant les décomptes de soins. Un règlement peut correspondre à plusieurs décomptes.

2.2 Les catégories de soins santé

L'assurance santé est répartie en de nombreuses catégories qui donnent lieu à différents types de prise en charge. Le but de cette partie et de présenter les principales catégories de soins en sa*nt*é.

Les consultations chez un médecin de ville

Un médecin de ville est un professionnel de santé exerçant en cabinet. Ce n'est pas un médecin en milieu hospitalier. Un médecin de ville peut être un médecin généraliste ou spécialiste.

Les consultations chez un médecin de ville peuvent donner lieu à différents actes comme par exemple des honoraires de médecin généraliste ou de médecin spécialiste.

Les soins paramédicaux – laboratoire – radiologie

Cette catégorie regroupe 3 parties :

- Les consultations et actes effectués chez un auxiliaire médical hors hôpital ou clinique. Un Auxiliaire Médical est un soignant non médecin. Les masseurs-kinésithérapeutes, infirmières, orthophonistes, orthoptistes, sages-femmes, pédicures-podologues sont des auxiliaires médicaux.
- Les examens effectués dans un Centre d'Imagerie Médicale, un cabinet de radiologie ou d'échographie, un cabinet médical équipé, un cabinet dentaire, hors hôpital ou clinique. L'Imagerie Médicale regroupe les différentes techniques permettant de fournir une reproduction visuelle du corps humain sans devoir recourir à la chirurgie. La Radiographie, l'IRM, le Scanner et l'Echographie font partie des actes d'imagerie médicale.
- Les analyses médicales effectuées auprès d'un laboratoire hors hôpital ou clinique. Les laboratoires d'analyses médicales ou laboratoires de biologie médicale (LBM) effectuent et interprètent des analyses sur des liquides ou prélèvements humains, dans le but de caractériser ou de suivre une maladie.

Les cures

Cette partie regroupe les cures thermales pour lesquelles il y a une participation de l'Assurance Maladie. Une cure thermale est un traitement médical prescrit par un médecin (généraliste ou spécialiste) qui se déroule dans une station thermale sur trois semaines, au cours de laquelle le patient est traité pour son affection par les eaux minérales naturelles des sources thermales et par leurs produits dérivés (gaz, boues …).

Le transport

Cette partie regroupe les transports effectués auprès d'un transporteur agréé hors transport d'urgence (SMUR) et pris en charge par l'Assurance Maladie.

La pharmacie

Cette partie regroupe les différentes prestations pouvant être effectuées par un pharmacien.

Un pharmacien peut délivrer des médicaments (remboursables ou non) mais également des Préparations Magistrales (préparations médicamenteuses effectuées, en l'absence de spécialité pharmaceutique, par le pharmacien pour un patient précis). Les Honoraires de Dispensation (rémunération perçue par le pharmacien pour la délivrance du médicament) constituent également une partie des actes regroupés dans la catégorie pharmacie.

L'appareillage

L'appareillage regroupe les aides matérielles pouvant aider dans la vie de tous les jours ou pour certains traitements. En font partie :

- Les aides auditives,
- Les ortho-prothèses,
- Les prothèses mammaires.

L'optique

Il s'agit des dispositifs médicaux et des prestations relatives à l'Optique :

- Lunettes (verres et monture),
- Lentilles,
- Chirurgie Réfractive de l'œil.

Le dentaire

Le dentaire regroupe les consultations et actes effectués chez un dentiste, un orthodontiste ou un stomatologue.

Il peut s'agir par exemple de :

- Soins dentaires,
- Prothèses fixes,
- Bridges,
- Prothèses amovibles,
- Implantologie,
- Traitements d'orthodontie.

L'hospitalisation

L'hospitalisation est l'admission d'un patient dans un centre hospitalier (hôpital ou clinique). Dans cette catégorie peuvent également figurer les soins externes qui sont des soins ambulatoires dispensés par les praticiens d'un hôpital ou une clinique (consultations, actes d'imagerie, analyses ...) et facturés par ce dernier pour lesquels il n'y a pas d'hospitalisation.

Les médecines douces

Prestations pour lesquelles il n'y a pas de participation du régime obligatoire comme les actes :

- D'un ostéopathe,
- D'un diététicien,
- D'un nutritionniste,
- D'un podologue,
- …

La Médecine Douce se différencie de la médecine officiellement reconnue en employant d'autres formes thérapeutiques.

2.3 La sécurité sociale

La sécurité sociale a été créée en 1945. D'après l'ordonnance n° 45-2250 du 4 octobre 1945, texte fondateur de la sécurité sociale, elle consiste en « la garantie donnée à chacun qu'en toutes circonstances il disposera des moyens nécessaires pour assurer sa subsistance et celle de sa famille dans des conditions décentes ». Elle a pour but d'aider les individus à faire face à différents risques liés à la vie : vieillesse, incapacité, invalidité, chômage ou enfants à charge.

La sécurité sociale se répartit en 3 types de régimes :

- Le régime général couvre plus de 80% des français. Il s'adresse aux travailleurs salariés, aux travailleurs indépendants et à toute personne bénéficiant de droit au titre de la résidence.
- Le régime agricole concerne les exploitants et salariés agricoles.
- Les régimes spéciaux comme ceux de la SNCF ou de l'Assemblée nationale par exemple.

Ces différents régimes ont des prises en charge différentes, cependant, les régimes tendent à se rapprocher sur les montants versés et les modalités de versement.

La Sécurité Sociale en France comporte actuellement 5 branches pour le régime général.

La branche maladie couvre en partie les conséquences :

- D'une invalidité : versement d'une pension lorsque l'assuré a une invalidité qui réduit sa capacité de travail,
- De la maternité : prise en charge d'examens avant et après la naissance et indemnités journalières durant le congé maternité,
- Liées à un décès : capital en cas de décès,
- D'un arrêt de travail : versement d'indemnités journalières.

Elle est gérée par la Caisse Nationale d'Assurance Maladie (CNAM).

La branche famille est gérée par la Caisse Nationale des d'Allocations Familiales (CNAF). Elle contribue aux aides à la famille (allocations familiales, prime de naissance, …) et aux aides au logement (APL).

La branche retraite est gérée par la Caisse Nationale d'Assurance Vieillesse (CNAV). Elle assure le calcul et le paiement des rentes de retraite ainsi que du minimum vieillesse.

La branche recouvrement est chargée de collecter l'ensemble des cotisations et des contributions de Sécurité Sociale auprès des entreprises et des particuliers. Elle gère la trésorerie de chacune des branches de la Sécurité Sociale. Elle est gérée par l'Agence Centrale des Organismes de Sécurité Sociale (ACOSS) et par l'Union de Recouvrement des Cotisations de Sécurité Sociale et d'Allocations Familiales (URSSAF).

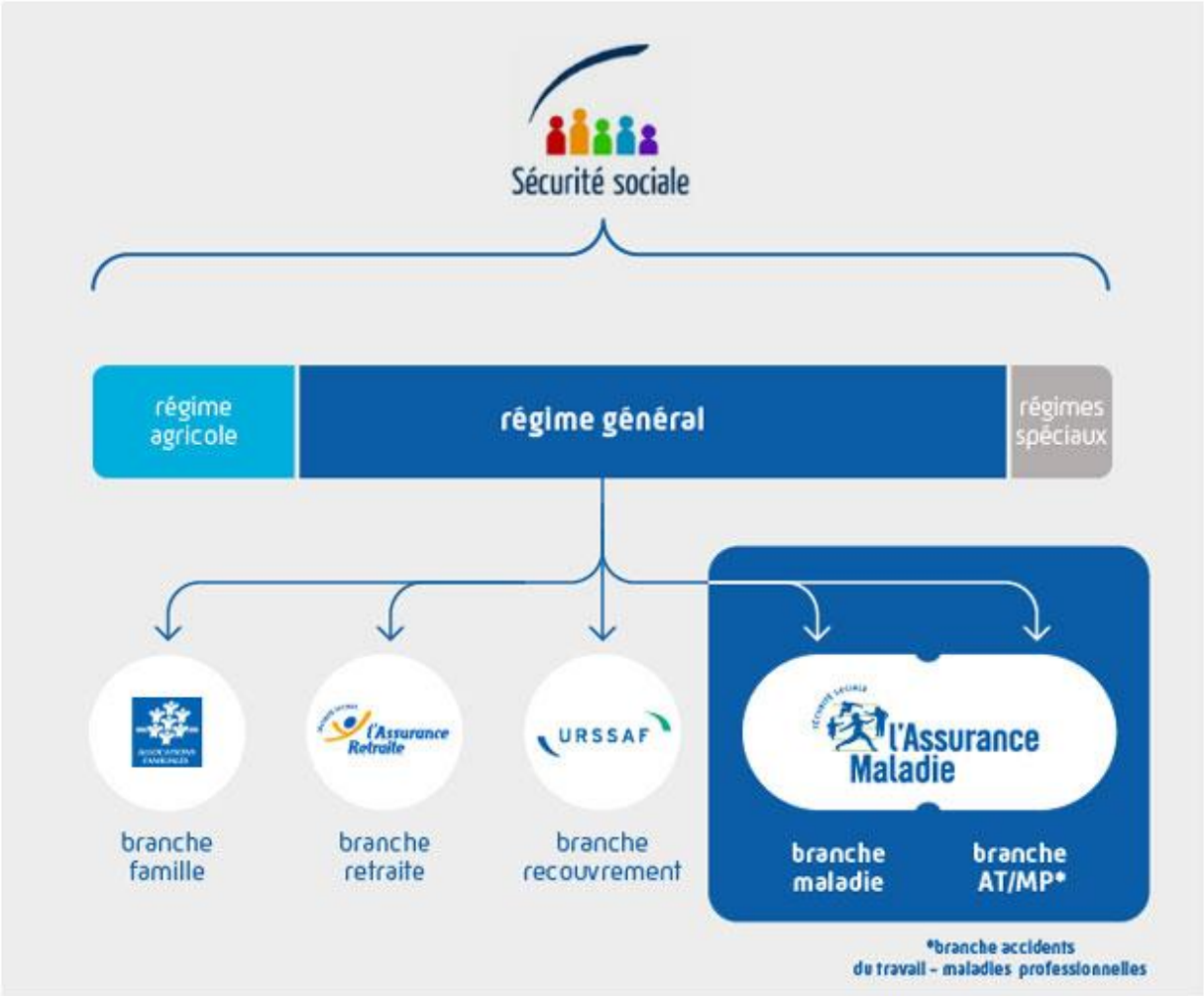


Figure 5 – L'organisation de la Sécurité Sociale[3]

Par une loi du 7 août 2020, la cinquième branche Autonomie a été créée. Elle est gérée par la Caisse Nationale de Solidarité pour l'Autonomie (CNSA) et a pour but de pour financer l'accompagnement de la perte d'autonomie des personnes âgées et des personnes handicapées.

2.4 Le remboursement de l'Assurance Maladie

L'Assurance Maladie correspond au premier niveau de prise en charge des dépenses de santé. Cette partie présente les différents éléments intervenant dans les remboursements de l'Assurance Maladie.

[3] **https://assurance-maladie.ameli.fr/qui-sommes-nous/organisation/securite-sociale**

- Les frais réels correspondent au montant demandé par le professionnel de santé pour le soin effectué.
- La base de remboursement de la Sécurité Sociale ou le tarif de responsabilité est la référence à partir de laquelle la Sécurité Sociale calcule son remboursement par acte de soins considéré.
- Le taux de remboursement correspond au pourcentage de la base de remboursement qui est pris en charge par la Sécurité Sociale.
- Le remboursement du régime obligatoire est égal à la base de remboursement multipliée par le taux de remboursement, c'est le montant que l'Assurance Maladie rembourse pour le soin considéré.
- Le ticket modérateur correspond à la part du tarif de responsabilité qui reste à charge de l'assuré.
- Le dépassement d'honoraires correspond aux frais réels imputés de la base de remboursement de la sécurité sociale.

2.5 La complémentaire santé

L'assurance maladie obligatoire prend en charge une partie des dépenses de santé. Pour se faire rembourser une partie ou tout le ticket modérateur, une complémentaire santé peut être souscrite. Une complémentaire santé est un contrat permettant de se couvrir contre les frais médicaux. L'assuré diminuera ainsi ses paiements directs qui sont parfois encore onéreux après remboursement de l'Assurance Maladie.

2.5.1 Les organismes d'assurance

Compagnies d'assurance

Les compagnies d'assurance sont séparées en deux groupes :

- Les sociétés anonymes sont des organismes de droit privé à but lucratif. Leur but est de réaliser des bénéfices pour les redistribuer à leurs actionnaires, les tarifs sont variables et dépendent de la situation de chaque adhérent.
- Les compagnies d'assurance mutuelle sont, d'après l'article L322-26-1 du Code des Assurances, des « personnes morales de droit privé ayant un objet non commercial ». Elles reposent sur des valeurs mutualistes mais ne doivent pas être confondues avec les mutuelles.

Les compagnies d'assurance sont régies par le Code des Assurances.

Mutuelles d'Assurances

D'après l'article L111-1 du Code de la mutualité, les mutuelles sont « des personnes morales de droit privé à but non lucratif. L'article L114-16 du même Code impose que les mutuelles soient dirigées par des personnes élues par les adhérents réunis en assemblée générale. Les mutuelles sont immatriculées au registre national des mutuelles.

Les mutuelles ne sont pas cotées en bourse : leur but n'est pas de redistribuer des bénéfices aux actionnaires. Le financement est effectué en grande majorité par les cotisations des membres.

Ces organismes reposent sur des valeurs de solidarité entre les adhérents, ils proposent des services comme :

- Des complémentaires santé,
- Des régimes de prévoyance,
- Des régimes supplémentaires de retraite.

Les cotisations ne dépendent pas des risques individuels propres à chaque adhérent.

Institution de Prévoyance

Les institutions de prévoyance sont des sociétés de personnes de droit privé. Elles sont régies par le Code de la Sécurité Sociale. Elles reposent sur un accord entre les partenaires sociaux et des branches professionnelles. Les conseils d'administration sont constitués à parts égales de salariés et d'employeurs décidant de la politique gestion des risques des collaborateurs.

Les Institutions de prévoyance sont à but non lucratif. Il n'y a pas de sélection des assurés. Leur activité principale est la couverture des risques de santé et de prévoyance pour les branches professionnelles. Elles peuvent aussi gérer les cotisations retraites de ces branches professionnelles, par délégation de la Sécurité sociale.

Les sociétés d'assurance, les mutuelles et les instituts de prévoyance peuvent proposer des complémentaires santé.

2.5.2 Les types de complémentaire santé

La souscription d'un contrat de complémentaire santé peut être faite :

- De manière individuelle,
- Dans le cadre d'un contrat collectif,
- Dans le cadre d'un contrat de complémentaire santé solidaire.

Le contrat individuel s'établit entre l'assureur et le souscripteur (et les autres assurés). Le tarif dépend alors des différents assurés du contrat.

Le contrat collectif est souscrit par une personne morale ou une entreprise. Il a pour but de couvrir les salariés. Le fait de mutualiser les risques au sein de l'entreprise permet alors des tarifs plus compétitifs. D'après les articles L911-1 à L911-8 du Code de la Sécurité Sociale, depuis le 1er janvier 2016, les employeurs du secteur privé (entreprise et association) ont obligation de proposer une complémentaire de santé collective à ses salariés et de participer à au moins 50% de la cotisation. Les articles D911-0 à D911-8 définissent le contenu de la mutuelle obligatoire.

Le contrat de complémentaire santé solidaire remplace la couverture maladie universelle complémentaire (CMUC). Il a pour but d'aider les personnes à faibles revenus face aux dépenses de santé. Il est gratuit ou payant en fonction des revenus.

2.5.3 Le remboursement de la complémentaire santé

La complémentaire santé intervient après l'Assurance Maladie. Cette partie présente les différents éléments intervenant dans les remboursements de la complémentaire santé.

- Le remboursement du régime complémentaire est le montant que rembourse la complémentaire santé. Ce montant peut prendre différentes formes :
    - Un pourcentage de la base de remboursement,
    - Un pourcentage des frais réels,
    - Un montant forfaitaire qui peut être un montant maximal par prestation ou un montant maximal par semestre ou par an, dans ce cas, il pourra être un pourcentage du Plafond Mensuel de la Sécurité sociale (PMSS),
    - Un pourcentage du Ticket Modérateur.
- Le reste à charge est le montant restant à payer pour le soin par l'assuré après déduction des remboursements du régime obligatoire et du régime complémentaire.

Le remboursement est synthétisé sur le graphique suivant :

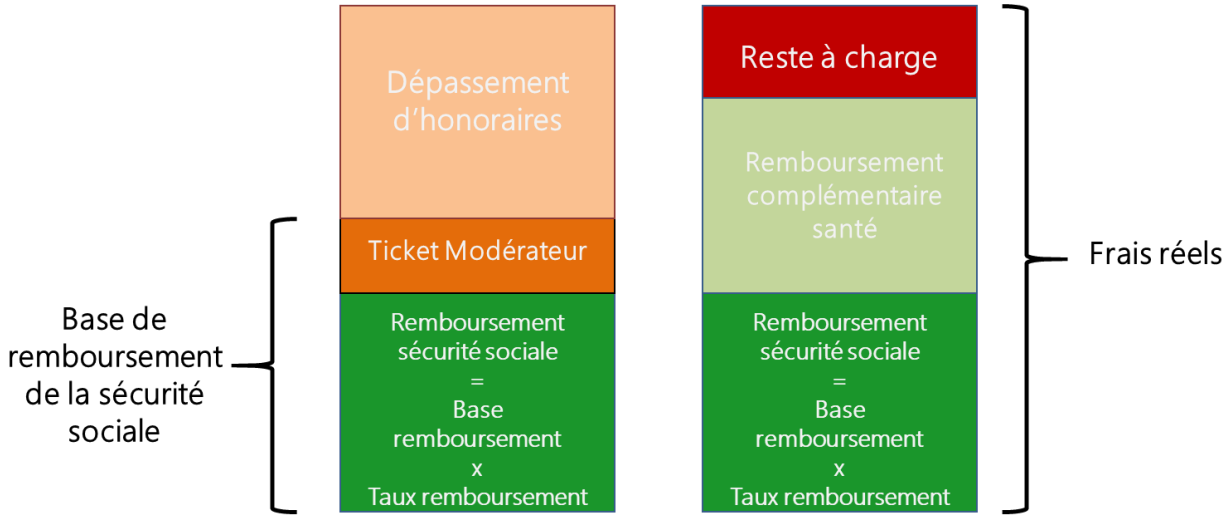


Figure 6 – Le remboursement d'un soin[4]

2.6 L'origine du décompte santé

Le traitement du décompte santé peut être réparti en 3 catégories :

- Le traitement des télétransmissions par régime obligatoire ou CETIP (tiers payant),
- Les documents lus automatiquement (lecture automatique des documents ou OCR),
- Les documents saisis au siège, ou télétransmis par internet ou par smartphone.

Le tableau ci-dessous regroupe les différents indicateurs dans les bases de données indiquant d'où provient le document et comment il a été traité :

Libellé de l' Emetteur d'origine du Décompte santé
TELETRANSMISSION REGIME OBLIGATOIRE
TELETRANSMISSION CETIP
SAISIE AU SIEGE (SAISIE AVANCEE)
TELETRANSMISSION Internet
TELETRANSMISSION Smartphone
TELETRANSMISSION OCR
LECTURE AUTOMATIQUE DES DECOMPTES

Figure 7 – Différentes voies de provenance des justificatifs de soins

[4] Inspiré de Maëva PALIS Titre Impact de la réforme du « 100 % Santé » sur les contrats individuels et collectifs de complémentaires santé

- La télétransmission permet à la plupart des organismes Sécurité sociale de transmettre informatiquement les informations nécessaires pour pouvoir rembourser l'assuré. Elle est active dès que l'adhésion est en cours.
- Le système de tiers payant généralisé permet à tout assuré de bénéficier de la "dispense d'avance des frais" prescrits médicalement et remboursables par leur régime obligatoire ou personnel d'assurance maladie. Ce système se nomme SP Santé. Le CETIP vérifie les formules, règle le professionnel de santé et envoie les renseignements à l'assureur pour remboursement. L'assureur rembourse le CETIP.
- Les documents saisis aux sièges, télétransmission par internet par l'assuré ou télétransmission par smartphone par l'assuré vont voir une plus grande intervention de la part des gestionnaires. En effet, ces derniers vont devoir analyser le document et extraire les différentes informations nécessaires au remboursement de l'assuré.
- La lecture automatique des documents est un ensemble d'outils permettant de récupérer les informations contenues dans le document. Cette technologie permet de reconnaître le type de document et d'extraire les informations nécessaires sans intervention d'un gestionnaire. L'*Optical Character Recognition* (OCR) est utilisé dans le même but.

2.7 La carte avance santé au Crédit Mutuel

La carte Avance Santé est une carte de paiement qui permet l'avance des frais de santé.

L'assuré peut payer avec la carte avance santé et n'avancera ainsi pas les frais de santé. Cette carte est liée au compte bancaire de l'assuré. L'assuré est débité sur son compte bancaire du reste à charge après intervention du régime obligatoire et de la complémentaire dans une limite de 30 jours.

Cette carte à plusieurs avantages :

- Elle permet de garantir la solvabilité de l'assuré envers le professionnel.
- Elle évite le risque que l'attestation de tiers payant ne soit pas à jour.
- La gestion administrative est plus simple lors du paiement, elle supprime les procédures de rejet dus à des contestations de tiers payant habituels.
- Les professionnels de santé sont crédités plus vite qu'avec un moyen de paiement classique.
- Elle permet également d'avancer les frais de médicaments non remboursables en pharmacie.

L'utilisation de la carte avance santé est renseignée dans les bases de données. Elle est un bon indicateur pour détecter les remboursements à tort. Par exemple, lors d'une suspicion de faux document, si l'assuré a utilisé la carte avance santé, il y aura moins de chances qu'il ait fraudé.

Les remboursements à tort en assurance santé peuvent désormais être présentés.

3. Les remboursements à tort en assurance santé

Le but de cette partie et de présenter les différents types de remboursements à tort en assurance santé, une méthode d'estimation de la fraude santé et quelques exemples de détection de remboursements à tort.

3.1 Types de remboursements à tort en assurance santé

En assurance santé, les remboursements à tort correspondent généralement aux catégories suivantes :

- Fraude de l'assuré,
- Fraude du professionnel de santé,
- Problème à la gestion,
- Recours potentiel.

La fraude des professionnels de santé ou des établissements de santé sont variés :

- Prestations fictives,
- Facturations multiples,
- Auto délivrance d'ordonnances,
- Facturation non conforme à l'acte réalisé,
- Optimisation de tarif pour coller au remboursement de la complémentaire santé,
- Surfacturation.

La fraude d'un assuré prend généralement une des formes suivantes :

- Falsification de documents existants (bénéficiaire, date, montant …)
- Création de toute pièce de documents,
- Usurpation ou faux papiers d'identités,
- Cumul de contrats santé,
- Utilisation frauduleuse de la carte vitale,
- Faux justificatifs de revenus ou de situation familiale.

Les autres indus pouvant intervenir lors de la gestion ont différentes causes :

- Règlements multiples pour un même soin,
- Remboursement de soins hors période de garantie,
- Mauvaise application de la garantie.

Différents recours peuvent également être effectués, pouvant permettre de récupérer des indus dans différents cas :

- Soins liés à un accident automobile non responsable (lunettes cassées, hospitalisation, …),
- Soins liés à une hospitalisation de plus de 30 jours,
- Soins liés à une ALD.

3.2 Modélisation de la fraude en assurance santé

Le but de cette section est de proposer une approche simple pour estimer la fraude en assurance santé.

Il est difficile d'estimer le montant de fraude. D'après la Tribune de l'Assurance, la fraude peut être évaluée à 7% des prestations versées chaque année[5]. Avec un benchmark, des taux de fraude minimum et maximum ont été obtenus par catégorie de prestation. Il est difficile de savoir le degré d'exactitude de ces chiffres mais le but de cette partie est de proposer une méthode d'estimation de la fraude à l'aide de ces éléments.

Le montant des prestations versées par catégorie de prestation a été récupéré. A partir des prestations versées et des taux de fraude minimum et maximum estimés, il est possible d'obtenir le montant minimum et maximum de fraude par catégorie de prestation et au global. Par souci de confidentialité, les chiffres ne sont pas présentés dans cette partie.

Soient :

- S_i le montant des prestations pour la catégorie de prestation i,
- $\tau_{min,i}$ le taux minimum de fraude pour la catégorie de prestation i,
- $\tau_{max,i}$ le taux maximum de fraude pour la catégorie de prestation i.

[5] https://tribune-assurance.optionfinance.fr/lessentiel/le-potentiel-de-fraude-en-europe-est-estime-a-10-des-prestations-payees.html

Le tableau suivant est obtenu :

Catégorie de prestation	Taux de fraude mini	Taux de fraude maxi	Prestation versées	Montant fraude mini	Montant fraude maxi
Optique	$\tau_{min,Optique}$	$\tau_{max,Optique}$	$S_{Optique}$	$\tau_{min,Optique} \times S_{Optique}$	$\tau_{max,Optique} \times S_{Optique}$
Dentaire	$\tau_{min,Dentaire}$	$\tau_{max,Dentaire}$	$S_{Dentaire}$	$\tau_{min,Dentaire} \times S_{Dentaire}$	$\tau_{max,Dentaire} \times S_{Dentaire}$
Hospitalisation	$\tau_{min,Hospi}$	$\tau_{max,Hospi}$	S_{Hospi}	$\tau_{min,Hospi} \times S_{Hospi}$	$\tau_{max,Hospi} \times S_{Hospi}$
Médecine douce	$\tau_{min,Méd\ douce}$	$\tau_{max,Méd\ douce}$	$S_{Méd\ douce}$	$\tau_{min,Méd\ douce} \times S_{Méd\ douce}$	$\tau_{max,Méd\ douce} \times S_{Méd\ douce}$
Transport	$\tau_{min,Transport}$	$\tau_{max,Transport}$	$S_{Transport}$	$\tau_{min,Transport} \times S_{Transport}$	$\tau_{max,Transport} \times S_{Transport}$
Kinésithérapie	$\tau_{min,Kiné}$	$\tau_{max,Kiné}$	$S_{Kiné}$	$\tau_{min,Kiné} \times S_{Kiné}$	$\tau_{max,Kiné} \times S_{Kiné}$
Audio prothèse	$\tau_{min,Audio}$	$\tau_{max,Audio}$	S_{Audio}	$\tau_{min,Audio} \times S_{Audio}$	$\tau_{max,Audio} \times S_{Audio}$

Figure 8 - Calcul des montants de fraude minimaux et maximaux par type de prestations

L'optique correspond à la catégorie de prestation pour laquelle le montant de fraude serait le plus important. L'hospitalisation arrive en deuxième du fait de son volume important de prestation. C'est d'ailleurs ces deux catégories de prestations qui semblent prioritaires dans les travaux de détection de la fraude chez les assureurs.

Soient :

- τ_i la variable aléatoire qui correspond au taux de fraude pour la catégorie de prestation i,
- $F_i = \tau_i \times S_i$ le montant de la fraude estimé pour la prestation i,
- $F = \sum_i F_i$ le montant total de la fraude estimée.

Pour modéliser le montant total de fraude, les hypothèses suivantes seront posées :

- Pour tout i, $\tau_i \sim \tau_{min,i} + (\tau_{max,i} - \tau_{min,i}) \times V_i$,
- Les V_i sont indépendantes identiquement distribuées de loi bornée sur $[0,1]$.

De ce fait, les τ_i sont indépendantes et bornées sur $[\tau_{min,i}, \tau_{max,i}]$.

Dans le cas où aucune hypothèse n'est faite sur le taux de fraude possible pour l'organisme d'assurance, une loi uniforme peut être utilisée.

Dans ce cas, V_i est supposée équiprobable dans l'intervalle $[0,1]$.

Sa fonction de densité est $f_{V_i} = 1_{[0,1]}$, son espérance est $\frac{1}{2}$ et sa variance $\frac{1}{12}$.

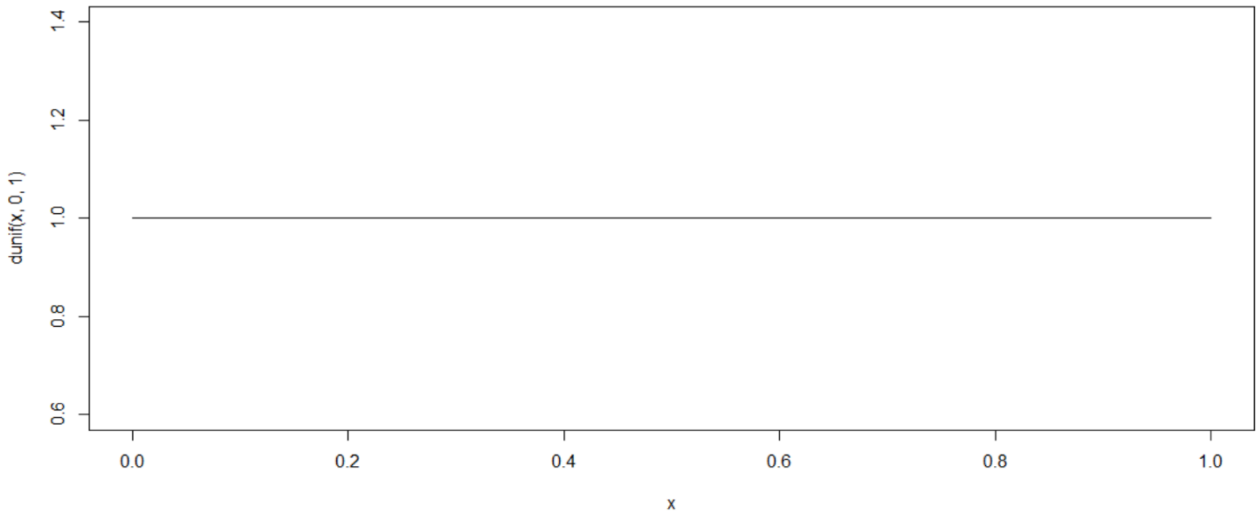


Figure 9 - Tracé de la densité de la loi uniforme à valeurs dans $[\mathbf{0}, \mathbf{1}]$.

Donc,

$E(F) = E(\sum_i F_i) = \sum_i E(F_i) = \sum_i \tau_{min,i} + (\tau_{max,i} - \tau_{min,i}) \times E(V_i) \times S_i$

Et finalement,

$$E(F) = \sum_i \tau_{min,i} + (\tau_{max,i} - \tau_{min,i}) \times \frac{1}{2} \times S_i$$

Si des hypothèses sont faites sur le taux de fraude, la loi bêta semble plus appropriée.

La loi bêta est définie sur $[0,1]$ et a deux paramètres notés α et β.

Sa fonction de densité est $f_{V_i} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\times\Gamma(\beta)} \times x^{\alpha-1} \times (1-x)^{\beta-1} \times 1_{[0,1]}$ où :

- Γ est la fonction gamma d'Euler,
- $\alpha > 0$,
- $\beta > 0$.

Son espérance est $\frac{\alpha}{\alpha+\beta}$ et sa variance $\frac{\alpha\times\beta}{(\alpha+\beta)^2\times(\alpha+\beta+1)}$.

Si $\alpha = \beta = 1$, la loi bêta est la loi uniforme.

En effet, $\frac{\Gamma(1+1)}{\Gamma(1)\times\Gamma(1)} \times x^{1-1} \times (1-x)^{1-1} \times 1_{[0,1]} = \frac{1!}{0!\times0!} \times x^0 \times (1-x)^0 \times 1_{[0,1]} = 1_{[0,1]}$.

Dans le cas, où il serait supposé que le taux de fraude en assurance santé aux ACM se situe environ dans la moyenne pour chaque catégorie de prestation, la loi bêta avec $\alpha = \beta$ peut être utilisée.

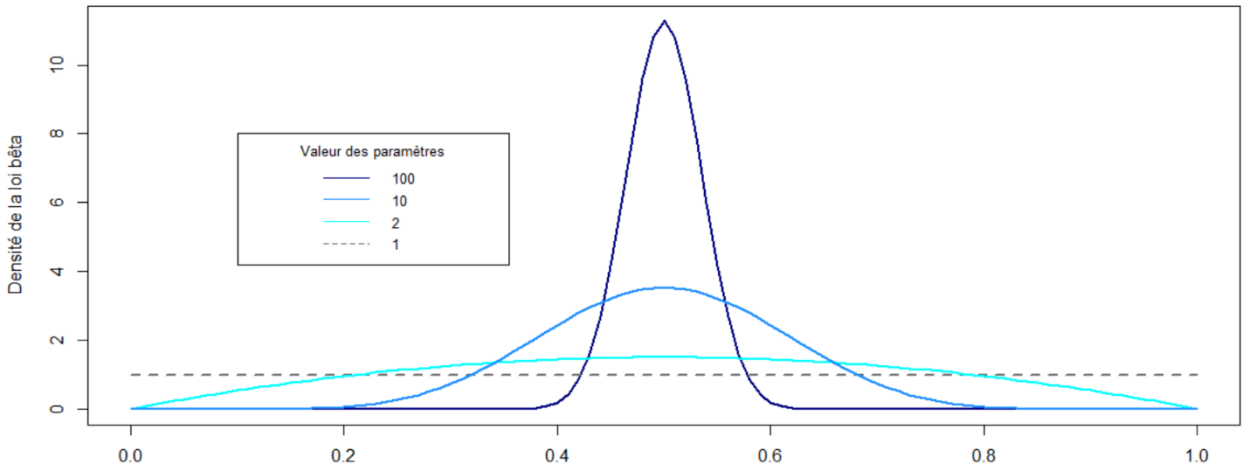


Figure 10 - Tracé de la densité de la loi $\boldsymbol{\beta}\hat{\boldsymbol{e}}\boldsymbol{ta}(\boldsymbol{\alpha},\boldsymbol{\beta})$ pour $\boldsymbol{\alpha = \beta \geq 1}$

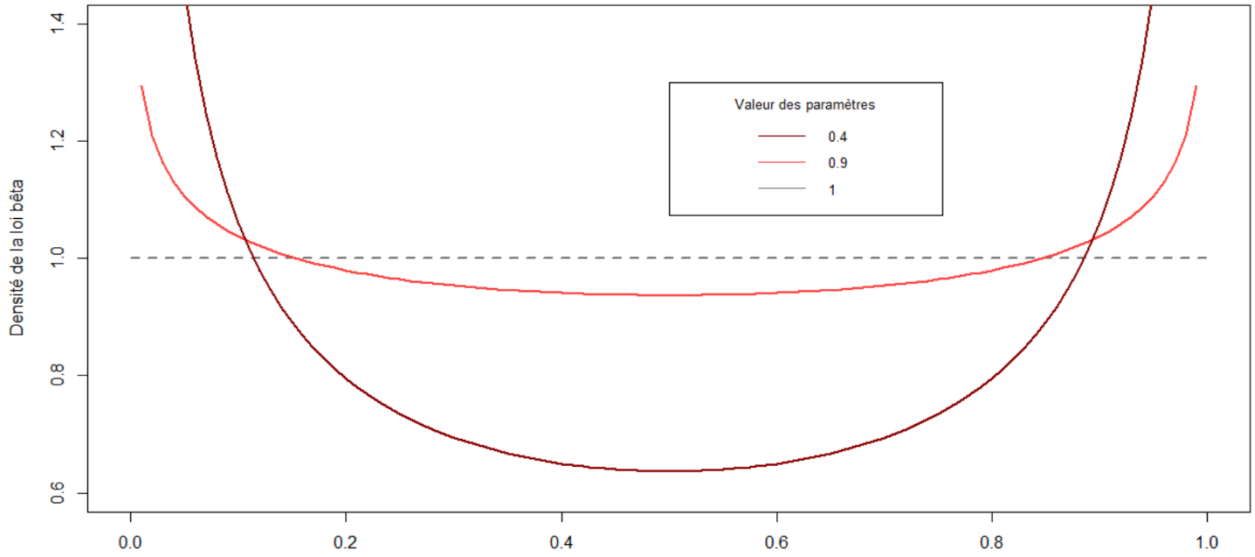


Figure 11 - Tracé de la densité de la loi $\boldsymbol{\beta}\hat{\boldsymbol{e}}\boldsymbol{ta}(\boldsymbol{\alpha},\boldsymbol{\beta})$ pour $\boldsymbol{\alpha = \beta \leq 1}$

Certaines valeurs de α peuvent être éliminées dans la modélisation. Les graphiques précédents permettent d'effectuer quelques conclusions :

- Pour $\alpha = \beta = 1$, la densité est celle de la loi uniforme.
- Pour $\alpha = \beta \leq 1$, la densité se représente sous une forme non cohérente avec les suppositions. En effet, pour qu'une telle forme soit utilisée, il faudrait que l'hypothèse de départ soit d'avoir le taux de fraude très faible ou très élevé par rapport au marché ce qui ne semble pas approprié.
- Pour $\alpha = \beta \geq 1$, la densité se représente sous une forme plus cohérente avec les suppositions. Plus $\alpha = \beta$ augmente, plus la variable aléatoire prend des valeurs proches de sa valeur moyenne et donc plus la $VaR_{99,5\%}$de la variable aléatoire diminue.

Si le taux de fraude est supposé plus faible que la moyenne du marché, il faut prendre $\alpha < \beta$ et ajuster les valeurs de $\alpha < \beta$ appropriées (plus α et β sont grands, plus la variable aléatoire est centrée sur l'estimation du taux de fraude sélectionnée). De même, si le taux de fraude est supposé plus élevé que la moyenne du marché, il faut prendre $\alpha > \beta$.

Une limite importante de ce modèle est qu'il est supposé que la répartition probable du taux de fraude soit identique pour toutes les prestations. Si les travaux de lutte contre la fraude sont inhomogènes selon les catégories de prestations, des lois bêta avec des paramètres différents peuvent être choisies pour les variables aléatoires modélisant les taux de fraude. Dans ce cas, ces variables aléatoires ne sont pas identiquement distribuées.

3.3 Requêtes effectuées dans le cadre des travaux afin de détecter des indus

3.3.1 Les chevauchements de périodes d'hospitalisation en chambre particulière

Pour analyser les séjours en hospitalisation, des prestations de chambres particulières ont été rapprochées.

Pour expliquer la démarche, différentes variables sont introduites :

- d_1 et f_1 les jours de début et de fin du premier séjour en chambre particulière,
- d_2 et f_2 les jours de début et de fin du deuxième séjour en chambre particulière,
- j_i le nombre de jours compris entre d_i et f_i,
- $j_i^{facturés}$ le nombre de jours facturés sur le séjour i,
- $P_i^{journalier}$ le prix journalier de la chambre particulière pour le séjour i.
- $P_i = j_i^{facturés} \times P_i^{journalier}$ les frais réels de la ligne de soins de chambre particulière du séjour i.

Le but est de trouver les facturations de chambres particulières pour des dates qui se chevauchent (exemple chambre particulière du 01 au 15 et du 02 au 16).

« Les séjours 1 et 2 ne se chevauchent pas » ⇔ «$f_1 < d_2$ » ou « $f_2 < d_1$ »

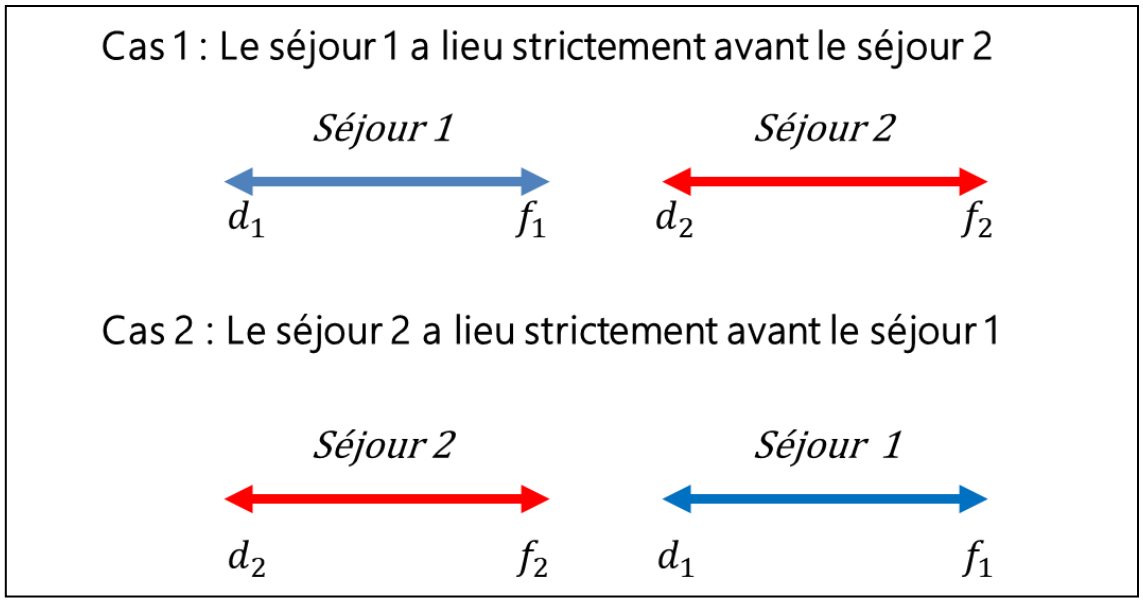


Figure 12 - Illustration des cas possibles où les séjours ne se chevauchent pas

Donc en utilisant les évènements contraires,

« Les séjours 1 et 2 se chevauchent » $\Leftrightarrow \overline{\text{« } f_1 < d_2 \text{ » } ou \text{ « } f_2 < d_1 \text{ »}}$

$\Leftrightarrow$ « $d_2 \leq f_1$ » et « $d_1 \leq f_2$ ».

Pour illustrer l'extraction, un cas est détaillé.

	Séjour 1	Séjour 2
Date de début de séjour	01/01/2022	10/01/2022
Date de fin de séjour	15/01/2022	30/01/2022
Nombre de jours du séjour	15	20
Nombre de jours facturés en chambre particulière	15	20
Frais réels de la chambre particulière	3000	3600
Prix journalier de la chambre particulière	200	180

Figure 13 – Exemple d'un chevauchement de séjours en chambre particulière

Le séjour 1 a lieu entre $d_1 = 01/01/2022$ et $f_1 = 15/01/2022$ donc sur une durée $j_1 = 15$ jours.

Le nombre de jours facturés est $j_1^{facturés} = 15$ jours.

Le prix total de la chambre particulière pendant le séjour est $P_1 = 3000€$ et donc $P_1^{journalier} = 200€$.

Le séjour 2 a lieu entre $d_2 = 10/01/2022$ et $f_2 = 30/01/2022$ donc sur une durée $j_1 = 20$ jours.

Le nombre de jours facturés est $j_2^{facturés} = 20$ jours.

Le prix total de la chambre particulière pendant le séjour est $P_2 = 3600€$ et donc $P_2^{journalier} = 180€$.

Le nombre de jours payés en trop est donc égal au nombre de jours entre le 10/01/2022 et le 15/01/2022 car ces jours ont été payés pour les deux séjours.

De manière générale,

$$Nombre\ de\ jours\ trop\ payés = Nombre\ de\ jours\ entre \max(d_1, d_2)\ et\ min(f_1, f_2).$$

D'après l'expérience des gestionnaires, dans ce type de cas, l'hôpital rembourse les soins correspondant à la chambre la moins chère facturée entre les deux séjours.

Donc ici, l'enjeu financier est $5 \times 180€ = 900€$.

De manière générale,

$$Enjeu\ financier = Nombre\ de\ jours\ trop\ payés \times \min\left(P_1^{journalier}, P_2^{journalier}\right).$$

Ces cas de soins payés en trop peuvent ne pas être détectés si seuls les séjours sur des dates identiques sont comparés.

En prenant en compte uniquement les chambres pour lesquelles le prix journalier est supérieur à 20 (pour éviter les erreurs de saisie et les mauvaises codifications d'actes), le tableau suivant est obtenu :

Acte de soins	Nombre d'observations	Moyenne	Ecart type	Minimum	Maximum
Supplément chambre particulière hospitalisation	444411	69	53	20,09	7520
Chambre particulière (acte 1)	28637	70	82	21	6090
Chambre particulière chirurgie	472	70	29	21	180
Chambre particulière (acte 2)	4	84	30	38,5	100
Chambre particulière ambulatoire	233	30	5	21	70
Supplément chambre particulière maternité	46015	71	53	20,4	2600

Figure 14 – Statistiques concernant les chambres particulières

Une extraction a été faite selon les critères suivants :

- Les actes de soins correspondent à ceux du tableau précédant,
- Les actes de soins ont été réalisés pour le même numéro de sécurité sociale et le même prénom,
- Il y a un chevauchement de dates de soins pour les périodes d'hospitalisation,
- Les dates de soins sont supérieures au 01/07/2017 (remontée à 5 ans),
- Il y a eu un règlement pour les soins et pas eu de dette ni d'annulation de règlement.

Après analyse de différents cas, certains types de faux semblants se distinguent.

- Les paiements de chambres particulières en deux étapes avec les jours facturés : $j_1 = j_2$ et $j_1^{facturés} + j_2^{facturés} = j_1$. Dans ces cas, les dates de soins correspondent à une période d'hospitalisation et la chambre particulière est réglée en deux étapes. De ce fait, sont retirés les lignes de soins pour lesquelles : $j_1^{facturés} + j_2^{facturés} \leq \max(j_1, j_2)$.
- Les paiements de chambre particulières en deux temps de la part de la gestion qui peuvent être dus à une erreur de saisie du remboursement ou à un remboursement sur deux contrats (complémentaire et surcomplémentaire) par exemple. C'est pourquoi sont conservées les lignes pour lesquelles la somme des remboursements des deux séjours dépasse les frais réels.
- Des autres cas de remboursements en deux temps détectés avec un prix journalier faible et un prix journalier plus élevé. Cela peut être dû à une erreur de saisie dans le prix de la chambre au départ qui donne un remboursement en deux temps. Par exemple, une chambre à prix journalier de 194€ et la deuxième rapprochée à 6€ par jour qui montre un remboursement en deux temps. C'est pourquoi seules les chambres pour lesquelles le montant journalier dépasse un certain seuil sont conservées.

Le but de cette étude est d'obtenir peu de faux-semblants pour faciliter le traitement des gestionnaires pour détecter d'abord des erreurs *a posteriori*. Une fois cette analyse, des solutions de détection *a priori* peuvent être mises en place. Les cas où les séjours se suivent et se chevauchent d'un jour ($f_1 = d_2$ ou $f_2 = d_1$) ont été écartés et traités à part. Les lignes pour lesquelles l'enjeu financier est au-dessus d'un certain seuil sont conservées.

À partir de l'analyse des cas sélectionnés, certaines typologies d'erreurs ressortent :

- Facture saisie deux fois sur un même contrat,
- Facture saisie deux fois sur deux contrats différents avec erreur du bénéficiaire réel du soin,
- Double facturation de l'établissement de santé pour la même période dans la même facture,
- Factures générées par l'établissement de santé (numéros de factures différents) avec des périodes de soins qui se chevauchent,
- Etablissement de santé qui génère une nouvelle facture suite à erreur initiale dans la première facture et pour lequel le remboursement a eu lieu deux fois.

3.3.2 Les soins hors période de couverture du contrat

Il se peut que des soins soient remboursés alors que l'assuré n'est plus couvert par son contrat.

Deux cas principaux sont possibles :

- Le contrat est clos mais l'assuré conserve et utilise son attestation de tiers payant. L'assureur a alors obligation de payer le praticien.
- La résiliation se fait dans certains cas à effet rétroactif. Par exemple, si un salarié d'une entreprise est couvert par un contrat collectif et quitte son entreprise, la date de fin de contrat pourra être celle de la fin de son contrat de travail. Or, il arrive souvent que cette information arrive tardivement. De ce fait, les paiements effectués concernant des soins ayant eu lieu pendant le différé doivent être remboursés par l'assuré. Cela peut également arriver pour des contrats qui sont annulés.

Une extraction a été effectuée pour obtenir les contrats pour lesquels :

- La date de fin est postérieure à une date fixée,
- Les codes produits ne correspondent pas à des surcomplémentaires,
- Il y a des soins sans régularisation avec une date de soins postérieure à la date de fin de contrat,
- Il n'y a pas de contrat en cours sur la période de ces derniers soins pour le même souscripteur.

Les montants potentiellement versés à tort ont été sommés et une analyse par type de résiliation a été effectuée (résiliation différée ou non, liquidation, contentieux).

Une extraction complémentaire a été effectuée pour détecter des paiements à tort sur deux contrats différents suite à une résiliation.

Les critères suivants ont été utilisés :

- Les soins ont été payés à la fois sur les deux contrats,
- Les soins du premier contrat ont une date de soins supérieure à la date de fin du contrat (soins hors période de garantie),
- Les frais réels, la date des soins, les bénéficiaires, le code acte sont identiques pour les deux lignes de soins des contrats,
- Il n'y a pas eu de régularisation pour les soins considérés,
- Il n'y a pas d'information de remboursement autre complémentaire sur la ligne de soins du nouveau contrat qui indiquerait la prise en compte de l'ancienne complémentaire dans le calcul du soin,
- Les deux contrats ont été souscrits par le même tiers.

Ces deux requêtes ont pour but de couvrir les remboursements à torts liés à une résiliation. L'idéal serait d'utiliser le numéro de sécurité sociale et pas le tiers souscripteur pour être parfaitement couvrant, pour n'avoir aucun faux semblant. Cependant, le numéro de tiers payant n'est pas toujours renseigné et cela implique un programme trop long à tourner sur tout l'historique des prestations.

À partir de ces requêtes, différents types d'indus ont pu être détectés. Des correctifs seront mis en place pour éviter certaines erreurs de gestion.

3.3.3 Les lignes de soins pour lesquels l'assuré percevrait plus que les frais réels

Les lignes de soins pour lesquelles $Frais\ réels < Montant\ CPAM + Montant\ ACM$ ont été étudiées. Deux groupes principaux ont été décelés parmi cette extraction :

- Des lignes pour lesquelles le remboursement du régime obligatoire est renseigné alors qu'il n'est pas réellement versé.
- Des lignes pour lesquelles le montant remboursé est environ égal à deux fois les frais réels.
- Des lignes correspondant à des régularisations de soins.

Pour les lignes présentant une erreur sur les remboursements du régime obligatoire, la quasi-totalité concerne des garanties d'orthodontie. Pour ces cas, le remboursement est un multiple du montant de la base de remboursement du régime obligatoire. Même si celle-ci n'est pas payée, elle doit être renseignée pour que le calcul du montant du remboursement soit bien effectué.

Pour les lignes pour lesquelles le remboursement ACM est bien supérieur aux frais réels, une partie de ces lignes correspond à un traitement du document par OCR sur de l'optique. En regardant quelques factures, le constat est que dans le calcul du remboursement ont été pris en compte deux fois les frais réels. Cette erreur n'est plus retrouvée depuis 2022, en effet, l'OCR est moins utilisé récemment sur ces types de soins et il a dû être amélioré avec le temps.

Une dernière partie des lignes remontées correspond en fait à des régularisations faites pour ajuster un autre remboursement.

Pour sélectionner les cas sur lesquels agir et éviter les faux-semblants, ont été conservés les cas où $Frais\ réels < Montant\ CPAM + Montant\ ACM$ au niveau global du décompte (qui regroupe plusieurs lignes de soins).

3.3.4 Les facturations incohérentes : exemple de l'honoraire de dispensation en pharmacie liée à l'âge de l'assuré

L'honoraire de dispensation est un frais que le pharmacien facture dans le cadre de la vente de médicaments remboursés par la Sécurité Sociale. Ces honoraires donnent des informations sur les médicaments. Il peut alors être intéressant d'étudier la cohérence entre les médicaments et les honoraires facturés. Les deux tableaux suivants montrent les différents honoraires de dispensation.

Médicament	Montant honoraire de dispensation		Taux de remboursement hors exonération du ticket modérateur			Taux de remboursement avec exonération du ticket modérateur
	« conditionnement normal »	« grand conditionnement »	Retraité et ayant droit	Agent en activité dans le cadre du libre choix	Agent en activité dans le cadre de la médecine de soins SNCF	
Service médical rendu faible - **PH2** (1)	**HD2** 1,02 euros	**HG2** 2,76 euros	75%	15%	100%	100%
Service médical rendu modéré - **PH4**	**HD4** 1,02 euros	**HG4** 2,76 euros		30%		
Service médical rendu important - **PH7**	**HD7** 1,02 euros	**HG7** 2,76 euros		65%		
Irremplaçable - **PH1**	**HD1** 1,02 euros	**HG1** 2,76 euros	100%	100%		

Taux et tarifs au 01/01/2020 hors DOM-TOM

Figure 15 – Les honoraires de dispensation par conditionnement[6]

Type d'ordonnance	Montant honoraire de dispensation	Taux de remboursement hors exonération du ticket modérateur			Taux de remboursement avec exonération du ticket modérateur
		Retraité et Ayant droit	Agent en activité dans le cadre du libre choix et les subsistants	Agent en activité dans le cadre de la médecine de soins SNCF	
Simple	**HDR** 0,51	75%	65%	100%	100%
Complexe (plus de cinq médicaments différents)	**HC** 0,31	100%	100%	100%	
En fonction de l'âge du patient (2)	**HDA** 1,58	75%	65%	Cas impossible	
Ayant au moins un médicament spécifique	**HDE** 3,57	75%	65%	100%	

Taux et tarifs au 01/09/2020 hors DOM-TOM

(2) Moins de 3 ans ou 70 ans et plus

Figure 16 – Les honoraires de dispensation dans la cadre d'une ordonnance[7]

[6] Source : **https://www.cprpsncf.fr/les-honoraires-de-dispensation-en-officine-details**

[7] Même source

L'honoraire HDA a été utilisée pour étudier la cohérence entre les soins et le bénéficiaire. En effet, cet honoraire dépend de l'âge du patient : il est facturé pour les patients de moins de 3 ans ou de 70 ans ou plus.

Ainsi, les lignes de soins pour lesquelles cet honoraire est facturé et l'âge du bénéficiaire au moment du soin est strictement compris entre 3 et 70 ans ont été regardées. Les cas remontés correspondent à des soins pour lesquels un bénéficiaire différent de celui de la prestation était renseigné. Des remboursements de soins pour des personnes non couvertes par le contrat ont été détectés (exemples d'enfants non couverts par le contrat qui sont alors sous le même numéro de sécurité sociale qu'un des deux parents).

Ces incohérences pourraient également permettre de faire remonter des fausses factures.

D'autres scénarios de facturations qui semblent incohérentes peuvent être analysés :

- La majoration enfant généraliste de 5 euros pour les enfants de moins de 6 ans pour une visite chez un médecin généraliste,
- Les verres progressifs pour des jeunes enfants.

3.4 Les soins incompatibles : la chirurgie réfractive suivie de soins optiques

La chirurgie réfractive est une opération visant à corriger les défauts visuels tels que la myopie, l'hypermétropie ou l'astigmatisme. Même si parfois la personne effectuant cette chirurgie doit encore porter des lunettes, sa vue est modifiée par la chirurgie. S'il achète des lunettes avant et après la chirurgie, les corrections doivent être différentes. La chirurgie peut être faite sur deux dates de soins différentes (une par œil) ou sur une date de soins.

Une base de données a été construite pour analyser les différentes atypies.

1. Sélection dans la table par ligne de soins des lignes pour lesquelles l'acte de soins est celui d'une chirurgie réfractive.
2. Jointure de la table obtenue avec la table des bénéficiaires.
3. Jointure avec la table des contrats.
4. Calcul de différents indicateurs (nombre de date de soins par bénéficiaire de chirurgie réfractive, durée entre la souscription et la chirurgie, âge du bénéficiaire au moment de la chirurgie …).

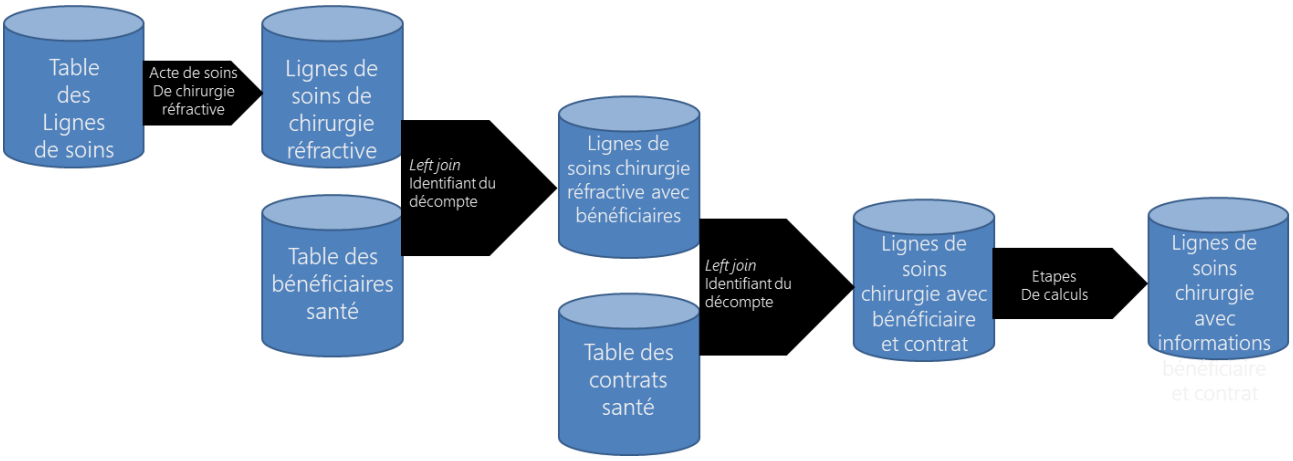


Figure 17 – Première étape de construction de la base des chirurgies réfractives

5. Sélection des bénéficiaires de chirurgies réfractives.
6. Obtention des soins optiques hors chirurgie réfractive de ces bénéficiaires.

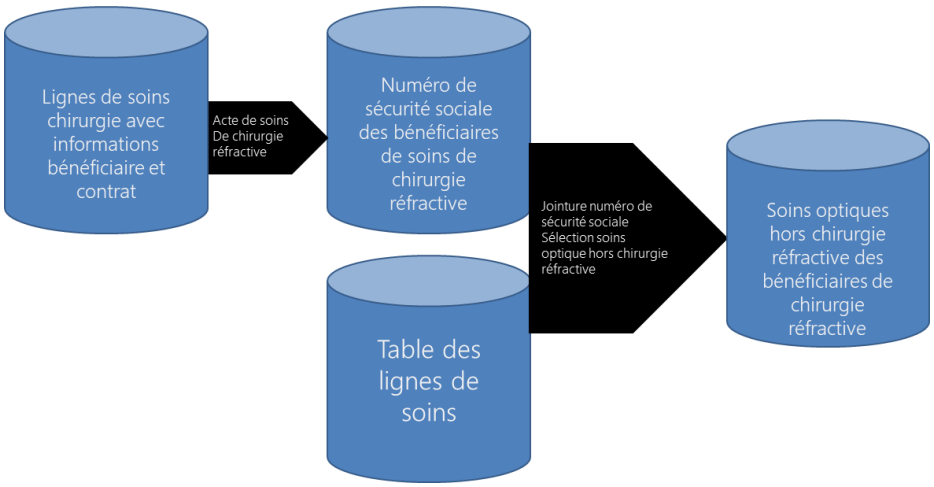


Figure 18 – Obtention des actes optiques hors chirurgie réfractive des bénéficiaires de chirurgie réfractive

7. Jointure à gauche de la table obtenue en étape 4 avec la table obtenue en étape 6.
8. Calcul de différents indicateurs par ligne de soins (concernant les soins optiques avant et après la chirurgie)
9. Sélection des variables souhaitées (présentées dans la partie suivante) et élimination des doublons

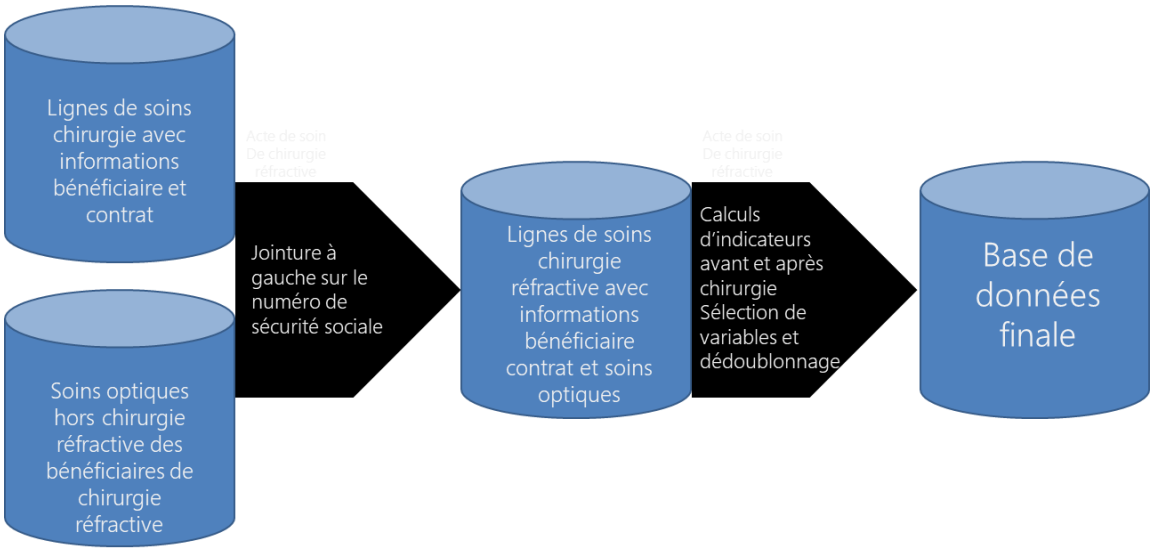


Figure 19 – Dernières étapes de construction de la base de données

Différents types de variables permettent alors d'analyser les prestations :

- La durée entre la souscription du contrat et la chirurgie,
- Le nombre de dates de soins de chirurgie réfractive pour le bénéficiaire,
- La somme des remboursements de chirurgie pour le bénéficiaire,
- L'âge du bénéficiaire au moment de la chirurgie,
- Le nombre de dates de soins optique distinctes avant et après chirurgie,
- Le nombre de verres simples avant et après chirurgie,
- Le nombre de verres complexes avant et après chirurgie,
- Le nombre de montures avant chirurgie,
- Le nombre de prestations lentilles avant et après chirurgie

À partir de cette base, différents graphiques ont pu être construits.

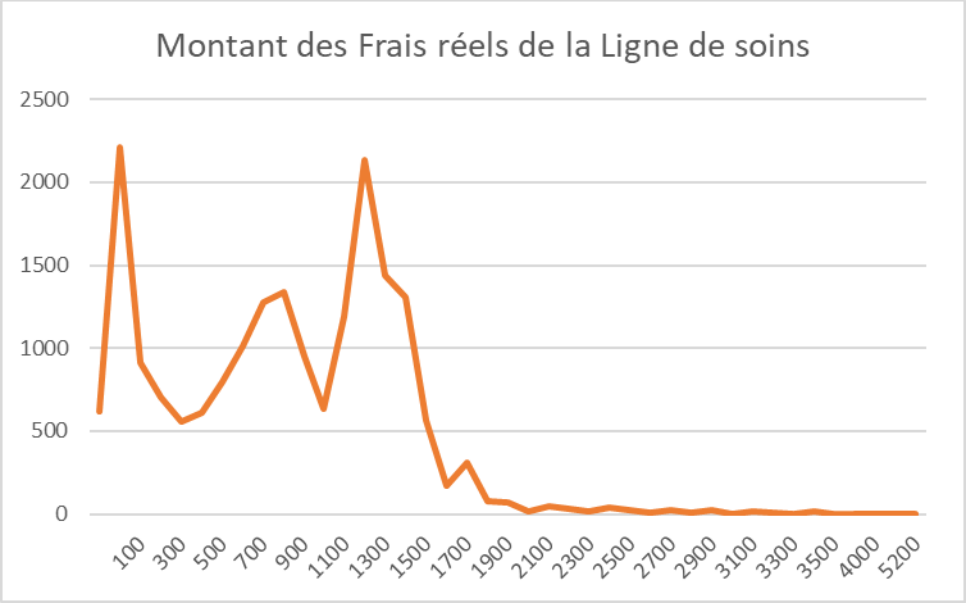


Figure 20 – Répartition des frais réels par ligne de soins

Deux groupes de frais réels sont retrouvés parmi l'acte de chirurgie réfractive. Une première partie concerne des montants faibles (inférieurs à 300€). Ces soins contiennent essentiellement des implants intraoculaires remplaçant le cristallin suite à une opération de la cataracte. Ils concernent ainsi des personnes en général de plus de 60 ans. Ces soins se repèrent par différents indicateurs :

- Des âges proches de 60 ans ou plus,
- Des montants plus faibles que les autres chirurgies réfractives,
- Un acte d'honoraires chirurgicaux le même jour lié à une opération de la cataracte.

D'autres prestations se retrouvent dans les montants faibles qui peuvent être des erreurs de codification d'acte ayant parfois des conséquences. Par exemple, le code de chirurgie réfractive et l'acte de kinésithérapeute sont proches. Une codification d'un acte de kinésithérapeute en chirurgie réfractive peut avoir des conséquences sur les forfaits annuels dont dispose l'assuré.

La deuxième partie des montants regroupe différentes catégories :

- Des frais liés à un œil ou deux suite à une chirurgie corrective de la vue (myopie, hypermétropie, astigmatie),
- Des honoraires chirurgicaux du médecin liés à cette chirurgie.

Ce sont ces derniers cas qui sont visés par l'analyse.

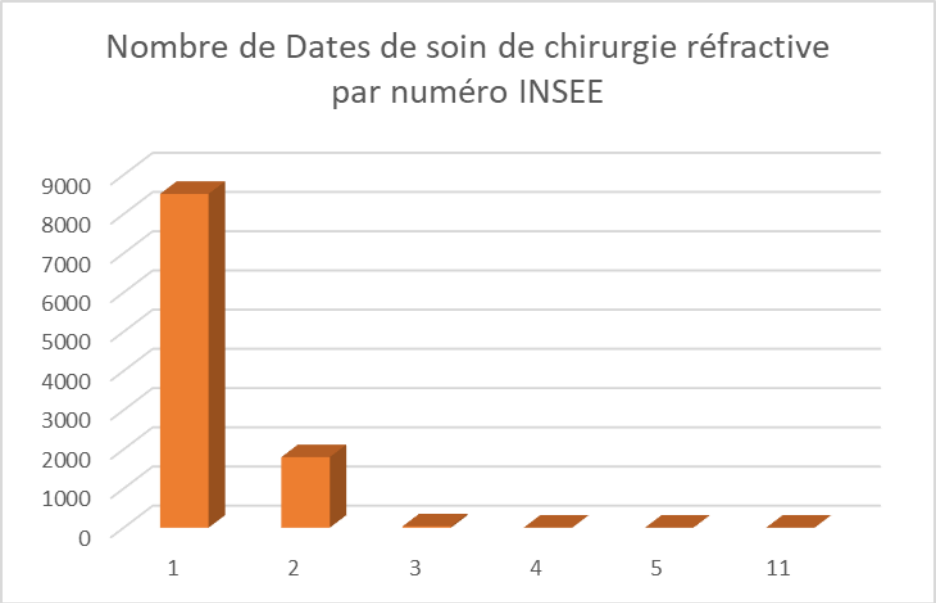


Figure 21 – Répartition du nombre de dates de soins de chirurgie réfractive par numéro de sécurité sociale

Une chirurgie des yeux est une opération à visée permanente. Une opération peut avoir lieu sur deux jours avec un jour par œil. Cependant, une répétition d'un acte chirurgie réfractive parait atypique et cela se confirme par les différents éléments de la base construite. Avec l'expertise du métier, différents cas de répétition de l'acte de chirurgie réfractive sont possibles. Par exemple, un patient peut se faire opérer de la myopie et, avec l'âge, développer une presbytie qui le conduira à effectuer une nouvelle opération des années plus tard pour corriger cette presbytie via une nouvelle chirurgie réfractive.

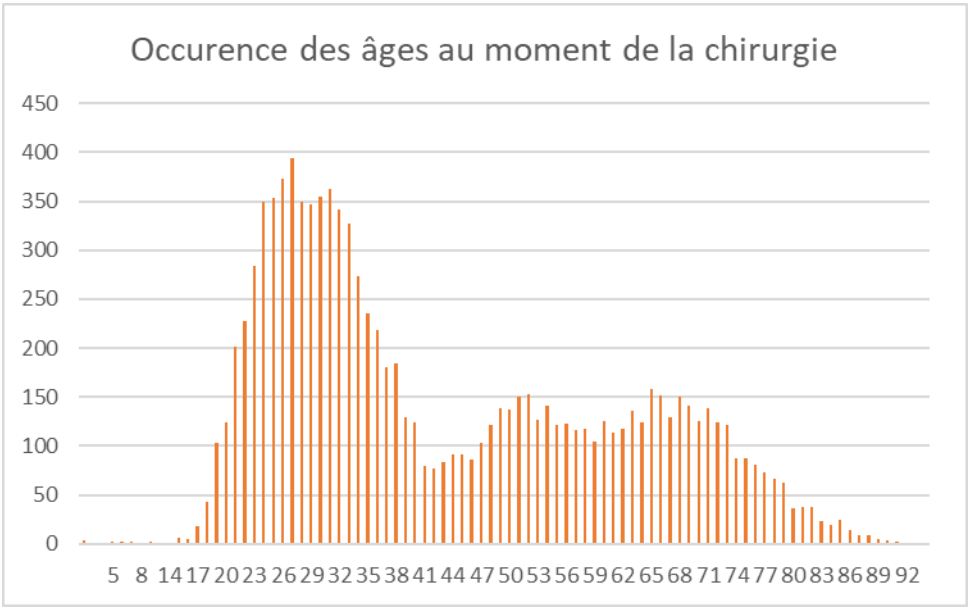


Figure 22 – Répartition des âges des assurés effectuant une chirurgie réfractive

La répartition des âges au moment de la chirurgie est en adéquation avec les attentes du métier. En effet, une partie importante des actes de chirurgie est effectuée pour corriger la myopie lorsque la vue est stabilisée et donc plutôt pour des patients entre 20 et 40 ans. Vers 45-50 ans, la presbytie peut se développer et conduire à une opération de chirurgie réfractive pour la corriger. Vers 60 ans, cette presbytie peut s'aggraver et les cas de chirurgie de la cataracte apparaissent.

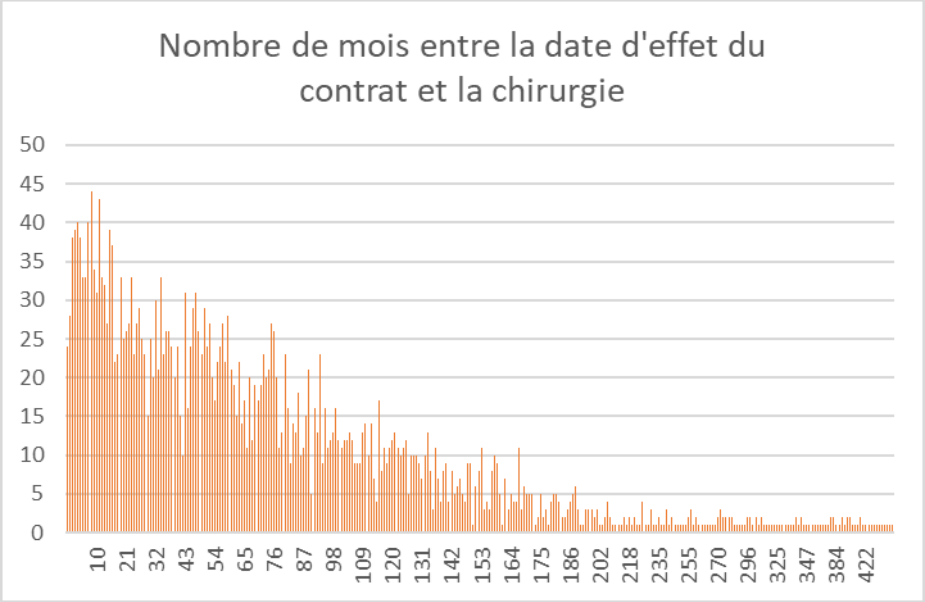


Figure 23 – Répartition du nombre de mois entre la date d'effet du contrat et la chirurgie

La chirurgie réfractive est souvent effectuée assez tôt dans la vie du contrat. D'après le graphique précédent, il peut même être constaté que la première année du contrat est celle où il y a le plus de chirurgie des yeux effectuée. Cela peut conduire à une perte de données pour exploiter des chirurgies fictives. En effet, si un patient effectue sa chirurgie quelques mois après la souscription du contrat, il sera difficile de savoir quelle était sa correction avant cette chirurgie. Il sera alors plus difficile de vérifier la cohérence des nouvelles prestations optiques après la chirurgie.

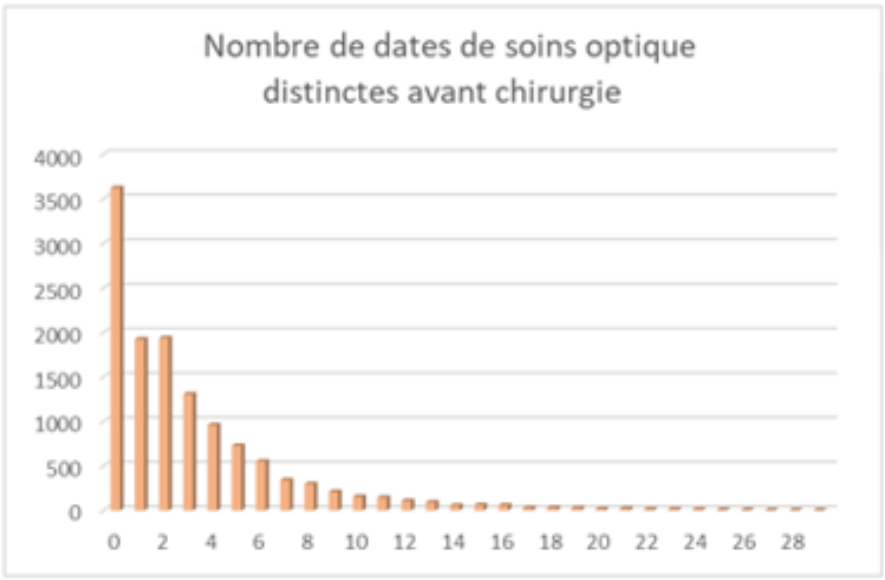


Figure 24 – Répartition du nombre de dates de soins optique distinctes avant la chirurgie réfractive

C'est donc parce que la chirurgie peut avoir lieu tôt dans la vie du contrat que pour une partie significative des cas il n'y a pas eu de prestations optiques antérieures à la chirurgie. C'est ce qui est visible sur le graphique précédant.

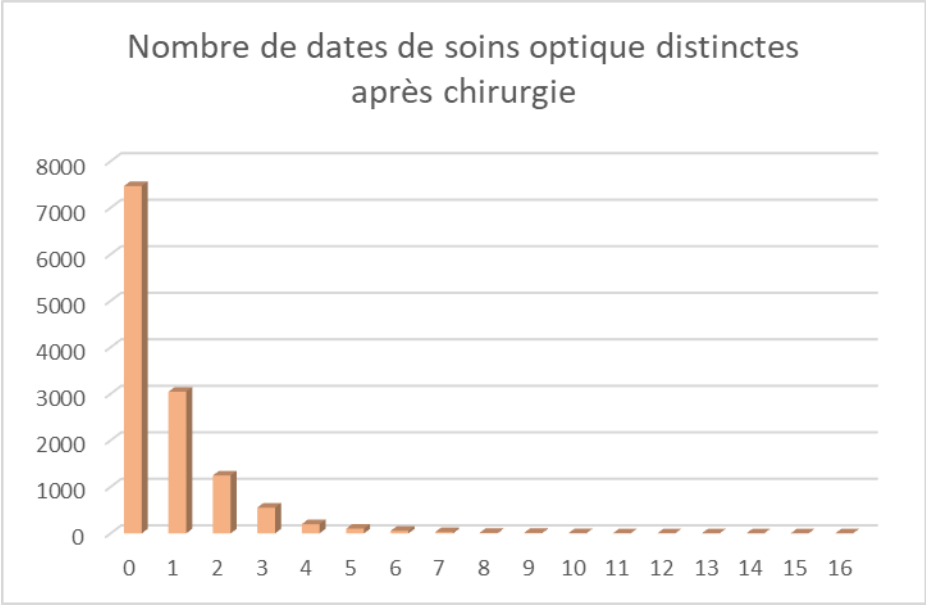


Figure 25 – Répartition du nombre de dates de soins optique distinctes après la chirurgie réfractive

Comme attendu, il y a néanmoins plus de personnes qui n'effectuent pas de prestations optiques après la chirurgie des yeux.

Les différentes prestations optiques (lentilles, verres simples, verres progressifs …) peuvent conduire à différents indicateurs indiquant des atypies. À partir des différentes études des cas extrêmes sur la base de données en termes d'âges, de montants, de prestations optiques, différents cas suspects ont pu être étudiées par le service anti-fraude et différents scénarios de fraude sur cet acte ont pu être établis. Des prestations réglées en double, des erreurs de gestion (bénéficiaire erroné, mauvais acte de soins) et des fraudes ont pu être décelées sur certains contrats. Cela a également permis de détecter des fraudes sur d'autres prestations sur ces contrats. Ainsi, des tests d'autres scénarios de fraude ont été effectués.

L'idéal serait, sur cet acte, d'utiliser différents algorithmes pour étudier les atypies. Dans ce cadre, différents algorithmes pourraient être utilisés comme proposés par Déborah HULOT dans son mémoire Implémentation d'un modèle de détection de fraude à l'assurance dans le cadre de soins hospitaliers. Ces algorithmes de détection des atypies pourraient également être appliqués à d'autres problématiques comme les soins optiques d'une famille ou les facturations de médicaments qui sont très encadrées en matière de prix et d'honoraires de dispensation.

Dans le même type d'idées, les éléments suivants pourraient également être analysés :

- Actes dentaires identiques répétés sur une même dent,
- Hospitalisation avec d'autres soins incompatibles pendant cette période de séjour.

Le choix pour ce mémoire a été d'étudier plus en détails les règlements en double.

4. Les doubles remboursements

Du fait du volume de prestations important en santé, il arrive que des prestations soient payées plusieurs fois par erreur. Ces règlements en double peuvent avoir plusieurs sources. Dans certains cas, la sécurité sociale rembourse elle-même l'assuré plusieurs fois ce qui peut induire l'assureur en erreur. Dans d'autres cas, un document transmis plusieurs fois par l'assuré ou une prestation traitée plusieurs fois par la gestion peut conduire à ces erreurs. C'est pourquoi des alertes sont mises en place pour lutter contre les doubles règlements. Cependant, certains règlements en double peuvent ne pas être détectés selon le calibrage de ces alertes qui doivent être à la fois pertinentes et gérables, c'est-à-dire en un nombre raisonnable pour pouvoir être traitées par l'équipe de gestion.

Le but de cette étude est de créer une base de données de doubles règlements potentiels en rapprochant des soins avec des caractéristiques proches. Une fois cette base construite, le mémoire se focalisera sur l'utilisation d'algorithmes d'apprentissage supervisé sur cette base de données. Les algorithmes permettront de détecter les doubles règlements *a posteriori* et des améliorations pourront être apportées aux alertes de doubles règlements *a priori* traitées par les gestionnaires.

4.1 Présentation du cadre de l'étude

4.1.1 Typologies de cas et d'indus

A l'aide de la construction de différentes bases de données différentes typologies de cas ont pu être détectés :

- Des télétransmissions passées deux fois pour une même prestation qui ont donné suite à un décompte négatif (acte de régularisation de la sécurité sociale suite à lequel l'assureur doit annuler un remboursement détaillé dans la prochaine sous-partie) qui n'avait pas été pris en compte,
- Double ou triple télétransmission d'un professionnel de santé sans régularisation de la sécurité sociale,
- Double règlement à un établissement de santé ou à un assuré suite à la saisie d'un même document deux fois,
- Règlement d'une même prestation sur deux contrats différents suite à une erreur de gestion,
- Règlement à l'assuré et également au professionnel de santé suite à utilisation du tiers payant et télétransmission par exemple.

Les doubles règlements peuvent être qualifiés de différentes façons en termes de types d'indus :

- S'il y a plusieurs télétransmissions d'un même soin par un professionnel de santé sans régularisation de la part de la Sécurité sociale, il peut s'agir d'une inattention mais également d'une fraude du professionnel.
- Si un document est télétransmis et envoyé par l'assuré avant tout remboursement, un double remboursement peut être une erreur de gestion. En revanche, si l'assuré envoie le document après avoir été remboursé suite à la télétransmission, il peut s'agir d'une mauvaise intention de sa part afin de se faire rembourser deux fois pour le même soin.
- Un flux de régularisation de la sécurité sociale non traité suite à un double règlement va être un problème de gestion.

4.1.2 Les régularisations de la Sécurité Sociale

Lorsque la Sécurité Sociale se rend compte qu'elle a transmis à tort un flux à l'assureur, elle émet un décompte négatif. C'est le cas notamment lorsque le professionnel de santé a facturé (par erreur) deux fois le même acte et qu'il s'en rend compte à postériori.

Le problème de ces flux est que l'assureur ne dispose pas de plus d'informations que les soins transmis à tort. L'assureur ne connait pas la raison de cette transmission à tort ce qui peut rendre difficile le traitement de ces décomptes négatifs.

Cette étude a montré que certains cas de décomptes négatifs dans le cadre de doubles règlements n'avaient pas été traités. Les cas non traités ont donc été étudiés et des solutions ont été mises en place.

4.1.3 Choix du cadre de l'étude

Dans le cas où un double règlement est détecté, une régularisation doit être faite. Une dette pouvant être des deux types suivants est créée :

- Une dette à valoir sur prestations : le montant se déduira des prochaines prestations payées à l'assuré.
- La dette à encaisser : le montant payé à tort est directement réclamé à l'assuré.

Si le double règlement est détecté assez tôt, le soin en doublon ne sera pas réglé ou le règlement effectué pourra être annulé si pas encore versé à l'assuré.

Plusieurs niveaux de détections peuvent être utilisés pour détecter les doubles règlements :

- La détection au niveau des règlements globaux de l'assureur,
- La détection au niveau des décomptes de soins,
- La détection au niveau des lignes de soins.

Le reste de cette partie a pour but de montrer que la détection est plus efficace au niveau des frais réels de la ligne de soins.

Premièrement, un règlement peut regrouper différents décomptes de soins. Ainsi, si un règlement ne regroupe pas les mêmes décomptes qu'un autre règlement, ces cas ne peuvent pas être retenus dans la détection des doubles règlements car non sélectionnés.

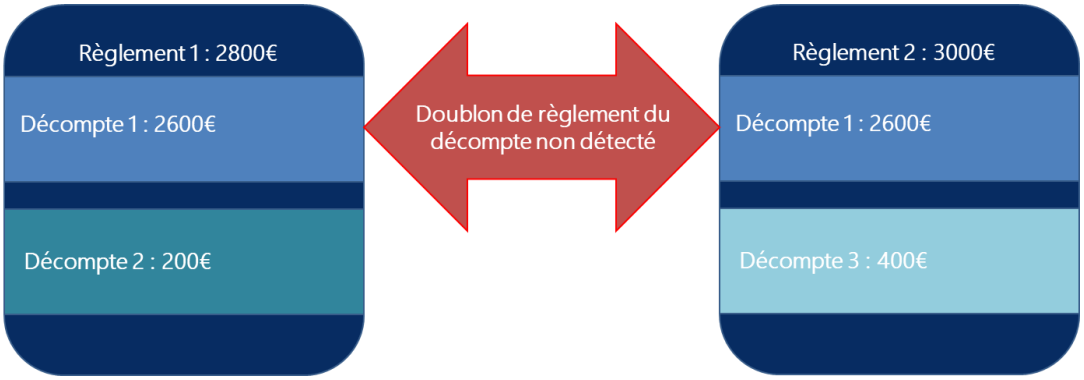


Figure 26 – Illustration du fait qu'il semble plus efficace de se placer au niveau du décompte que du règlement

De la même manière, si un soin donne lieu à deux lignes de soins réparties en deux décomptes, en se plaçant au niveau du décompte, le lien n'est pas fait si ces lignes de soins sont saisies sur un unique décompte. C'est également pour cela qu'il est plus pertinent de se placer au niveau de la ligne de soins pour éviter de perdre une partie des informations.

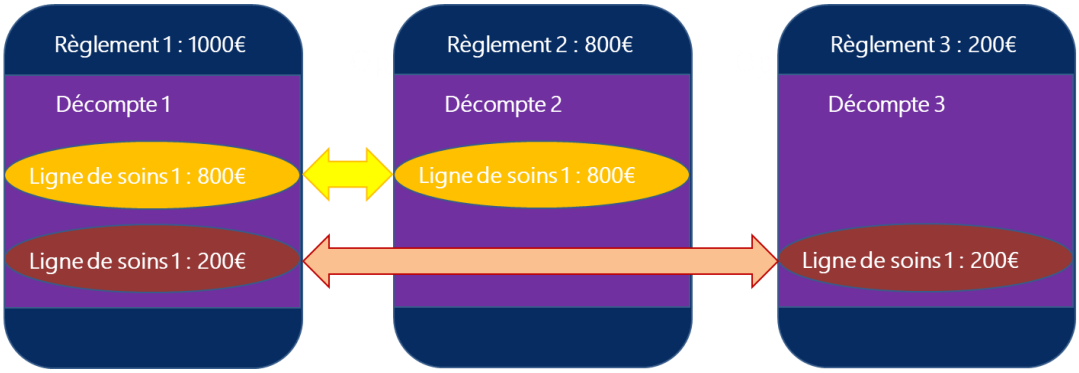


Figure 27 – Illustration montrant qu'il semble plus judicieux de se placer au niveau de la ligne de soins

Enfin, si deux règlements se font pour des soins identiques mais si le calcul des remboursements a été différent au niveau de la gestion (cela peut être dû à un problème au niveau du flux télétransmis par exemple une inversion entre le taux du régime local ou du régime général, ou à une erreur de saisie d'un montant de remboursement), alors l'information ne sera détectée qu'en se plaçant au niveau des frais réels de la ligne de soins.

Figure 28 – Illustration montrant qu'il semble plus judicieux de se placer au niveau des frais réels

Ainsi, les doubles règlements seront analysés au niveau de la ligne de soins pour éviter toute perte d'information.

4.2 Construction de la base de données

Une base de données a été construite pour détecter les cas où une dette n'a pas été mise en place (double de règlement sans régularisation). Une ligne contient les informations de deux lignes de soins potentiellement en doublon.

Les critères d'extractions sont les suivants :

- Les dates de soins sont identiques,
- Le montant des frais réels est identique,
- Les soins ont été réalisés pour le même bénéficiaire,
- Chaque ligne de soins a fait l'objet d'un règlement de plus de 50€,
- Les deux lignes de soins ont été renseignées sur deux décomptes différents,
- Les deux lignes de soins ont été effectuées sur un même contrat.

Après analyse, des critères ont permis de supprimer des faux-semblants évitables de la base de données :

- La prise en compte de l'oreille dans le cas d'audioprothèses quand elle est renseignée,
- La prise en compte du numéro de dent lorsqu'il est renseigné.

Ainsi, sont éliminés les cas pour lesquels :

- Dans le cas du dentaire, la dent est renseignée pour une ligne de soins et pas renseignée ou correspondant à une autre dent pour l'autre ligne de soins,
- Dans le cas des aides auditives, la position de l'oreille est renseignée pour une ligne de soins et pas renseignée ou correspondant à l'autre oreille pour la deuxième ligne de soins.

L'extraction est effectuée selon les étapes suivantes :

1. Jointure de la table des lignes de soins sur elle-même avec sélection des lignes pour lesquelles :
    - Les frais réels sont identiques,
    - Les dates de soins sont identiques,
    - Les identifiants des décomptes santé sont différents,
    - Les numéros de dents et oreilles renseignées sont identiques ou non présents.
2. Jointure à gauche de la table obtenue avec la table des dettes santé pour obtenir, s'il y en a un, un identifiant de dette santé sur le décompte correspondant. Sélection des lignes de soins pour lesquelles les identifiants de dette sont manquants (donc sélection des lignes de soins sans régularisation de type dette).
3. Jointure de la table obtenue avec la table des règlements santé et sélection des lignes de soins pour lesquelles il y a des règlements et pour lesquelles les règlements n'ont pas été rejetés.
4. Jointure de la table obtenue avec la table des bénéficiaires de décomptes santé et sélections des lignes pour lesquelles les deux lignes de soins correspondent au même bénéficiaire.

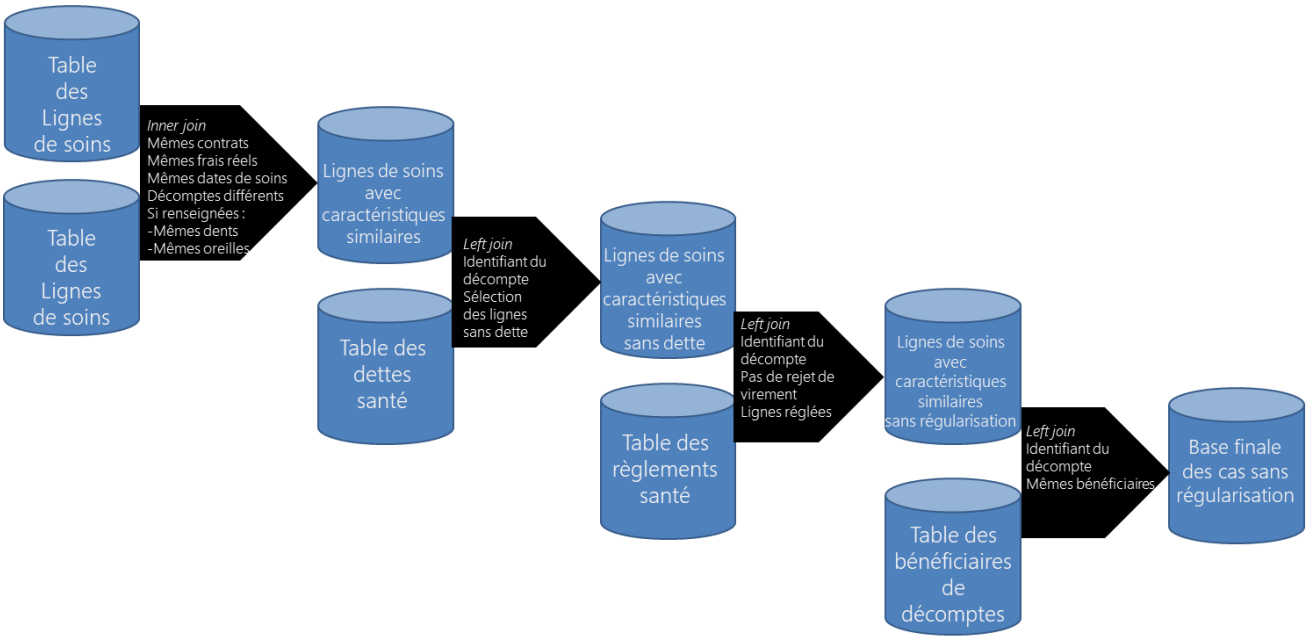


Figure 29 – Extraction des cas sans régularisation

La base est constituée des éléments pour lesquels il n'y a pas eu de régularisation. Une extraction similaire est effectuée pour récupérer tous les cas où il y a eu des régularisations. Ces cas ont donc une forte chance de correspondre à des doubles règlements.

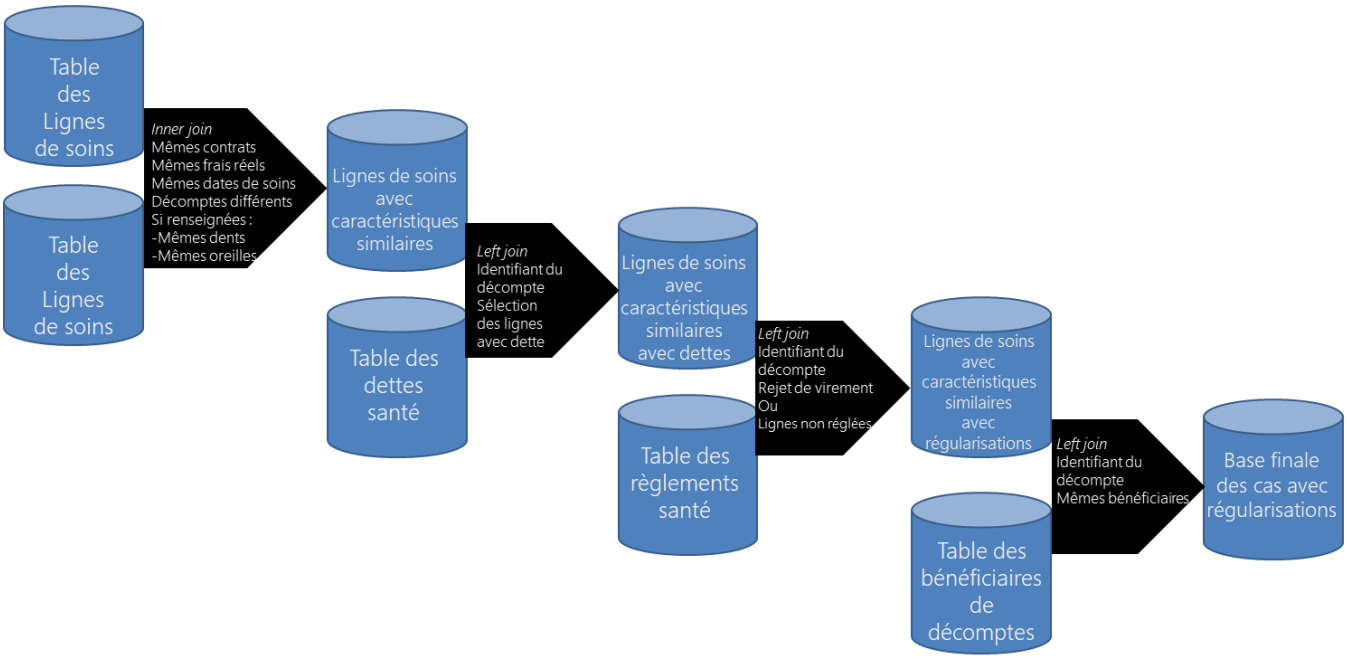


Figure 30 – Extraction des cas avec régularisation

La base finale est alors la concaténation des cas où il y a eu régularisation et des cas où il n'y a pas eu de régularisation. Ainsi, la base est constituée de 50% de lignes où il y a eu des régularisations dans la base de données.

4.3 Ajout d'indicateur à la base de données

Pour pouvoir analyser les lignes et pour calibrer les algorithmes de prédiction, différents indicateurs ont été ajoutés à la base de données. Ces indicateurs sont utilisés pour construire la base de données finale. Ils sont utilisés pour déterminer si les lignes de la base de données correspondent à un doublon.

4.3.1 Regroupement d'actes de soins

Différents actes de soins ont été regroupés. En effet, même si un même soin devrait toujours être saisi sous le même code acte, différents actes proches peuvent porter à confusion.

Ainsi, pour calibrer les algorithmes de prédiction, différents actes ont été regroupés. Voici quelques exemples de regroupement :

- Le Poste Dentaire est réparti en trois groupes nommés Orthodontie, Dentaire et Complément Dentaire,
- Un groupement Honoraires pour actes médicaux est constitué des honoraires chirurgicaux, des honoraires pour hospitalisation médicale, des honoraires maternité et actes médicaux,
- Un groupement Chambre Particulière correspond aux chambres particulières, aux suppléments chambre particulière hospitalisation et maternité,
- Les différents actes correspondant à des montures, des verres, des lentilles, sont regroupés en Monture, Verre et Lentille.

4.3.2 Indicateurs comparant les deux lignes de soins sélectionnées

Chaque ligne est composée d'indicateurs sur les deux lignes de soins.

Différents indicateurs ont été créés afin de les comparer pour transformer les montants de remboursements de chaque ligne de soins :

- Nb_ligne_reg_acte correspond au nombre de lignes extraites pour un même jour, contrat, bénéficiaire et regroupement d'actes de soins. Si ce chiffre est supérieur ou égal à 1 cela signifie que les actes des lignes de soins correspondent au même regroupement d'acte. Cet indicateur est important en optique. En effet, pour les verres, s'il y a doublon, comme une ligne de soins correspond à un verre, les deux

lignes de soins en doublon vont être reliées aux deux autres lignes de soins et quatre lignes seront extraites.

- Reg_GECOR_ass vaut 1 si une ligne de soins a fait l'objet d'un règlement à l'assuré et l'autre à un professionnel de santé et 0 sinon. Cela fait suite à l'observation qu'Il peut y avoir des doublons si un soin passe par télétransmission et un autre par tiers payant.
- FR_nuls vaut 1 si les frais réels sont nuls et 0 sinon.
- Mm_reg vaut 1 si les deux lignes de soins ont été payées dans un même règlement et 0 sinon.
- MM_acte vaut 1 si les deux lignes de soins sont saisies sous le même acte de soins et 0 sinon.
- Deux_tele vaut 1 si les deux lignes de soins ont été payées suite à une télétransmission du régime obligatoire et 0 sinon.
- Montants_proches_RO vaut 1 si les remboursements du régime obligatoire sont proches (1% d'écart) pour les deux lignes de soins et 0 sinon.
- Plus_perçu vaut 1 si :

$$Perçu = RembACM_1 + RembACM_2 + Max(MontantRO_1, MontantRO_2) > FraisRéels$$

  Et 0 sinon.
- Perçu_proche_FR vaut 1 si $Perçu$ et les Frais réels sont proches (2% d'écart) et 0 sinon. Si le montant perçu est proche des frais réels, cela peut être un indicateur de paiement en deux temps.
- Remb_proches vaut 1 si les deux remboursements ACM sont proches (1% d'écart). Si une même ligne de soins est payée deux fois et que le forfait est encore disponible pour les deux lignes, le montant remboursé devrait être le même.

4.4 Retraitement

L'indicateur Doublon est retraité à l'aide de règles métier. Ces règles sont différentes pour chaque type d'acte de soins. Pour certains actes, il est difficile de retraiter les lignes avec certitude, seules deux factures identiques ou un décompte négatif de la sécurité sociale permettrait de dire que la ligne de soins a été payée en double. Les règles métier ne sont pas présentées dans le mémoire.

Pour les lignes sélectionnées pour lesquelles il y a déjà eu une régularisation, cela ne signifie pas forcément qu'il y a eu un doublon de règlement. C'est pourquoi 4% de ces cas ont été considérés comme non doublons.

De même, pour les lignes sans régularisations, certains cas font suite à des doublons non régularisés. L'utilisation de règles métier a conduit à considérer 25% de ces cas comme doublons. Pour certaines lignes considérées non doublons, il est difficile de conclure. En effet, pour le dentaire où les dents ne sont pas renseignées, il faudrait regarder toutes les factures pour savoir s'il y a des règlements en doublon ou non.

Au final, la base de données est constituée de lignes considérées comme doublons pour 60,5% des cas.

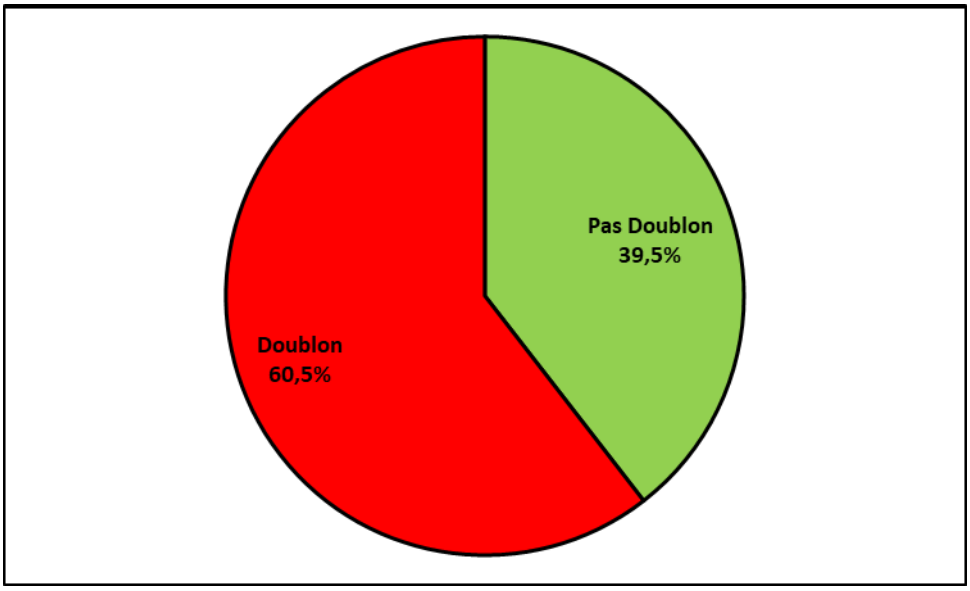


Figure 31 – Répartition de l'indicateur doublon dans la base de données retraitée

4.5 Analyse de la base de données retraitée

4.5.1 Les actes de soins

Les répartitions des doublons et des lignes de soins sont représentées à l'aide des deux graphiques suivants :

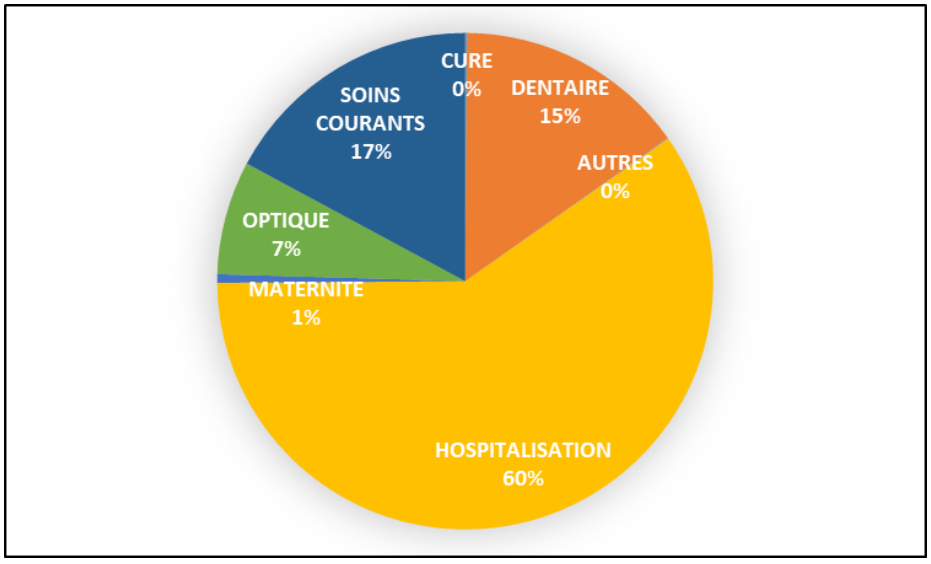


Figure 32 – Répartition des lignes avec doublon en fonction du secteur de soins

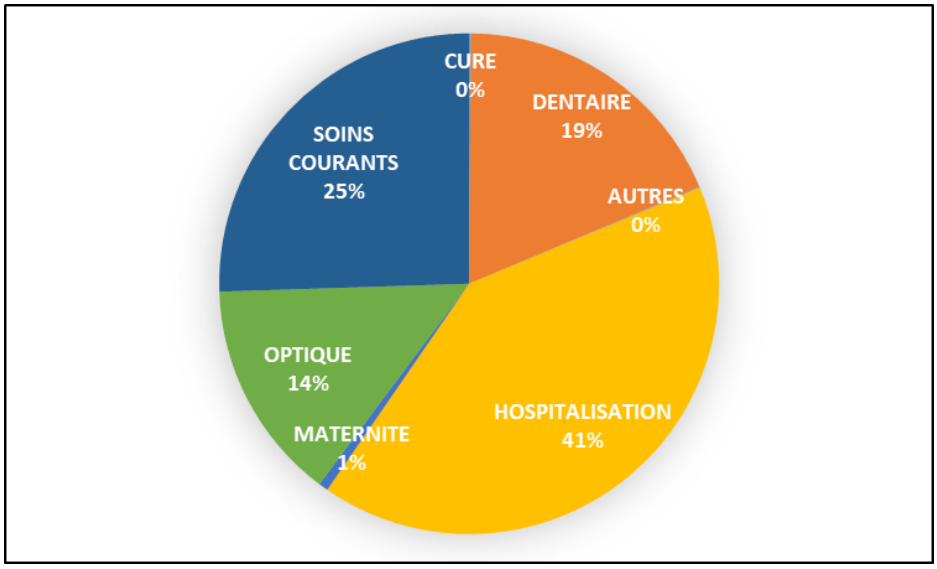


Figure 33 – Répartition des lignes extraites en fonction du secteur de soins

4.5.2 Représentation des indicateurs en fonction de la variable Doublon

La densité du nombre de ligne par regroupement d'acte peut être représentée en fonction de la colonne doublon.

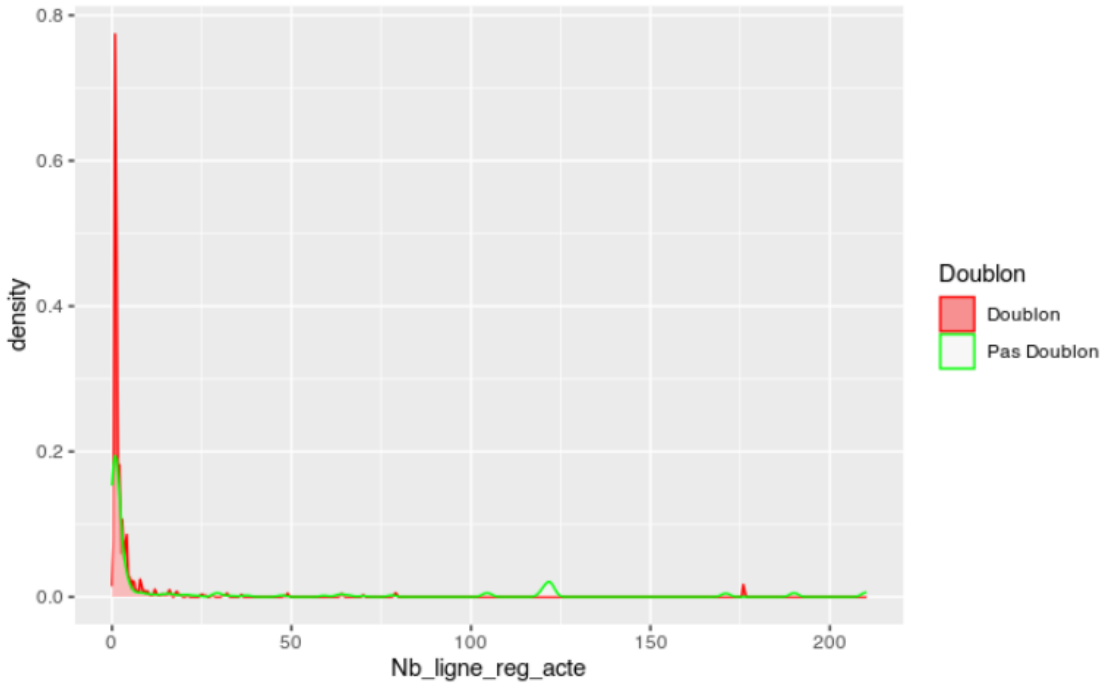


Figure 34 – Densité de la variable Nb_ligne_reg_acte sachant la variable Doublon

Environ 57% des doublons correspondent à Nb_ligne_reg_acte=1. Au-delà de Nb_line_reg_acte=50, il y a peu de doublons. Les cas pour lesquels Nb_ligne_reg_acte est supérieur à 50 correspondent en majeure partie à un établissement de santé qui facture tous ses soins à une même date de soins. Cela crée alors des atypies. De plus, Nb_ligne_reg_acte vaut 1 pour environ la moitié des lignes.

En catégorisant la variable Nb_ligne_reg_acte, le graphique suivant est obtenu.

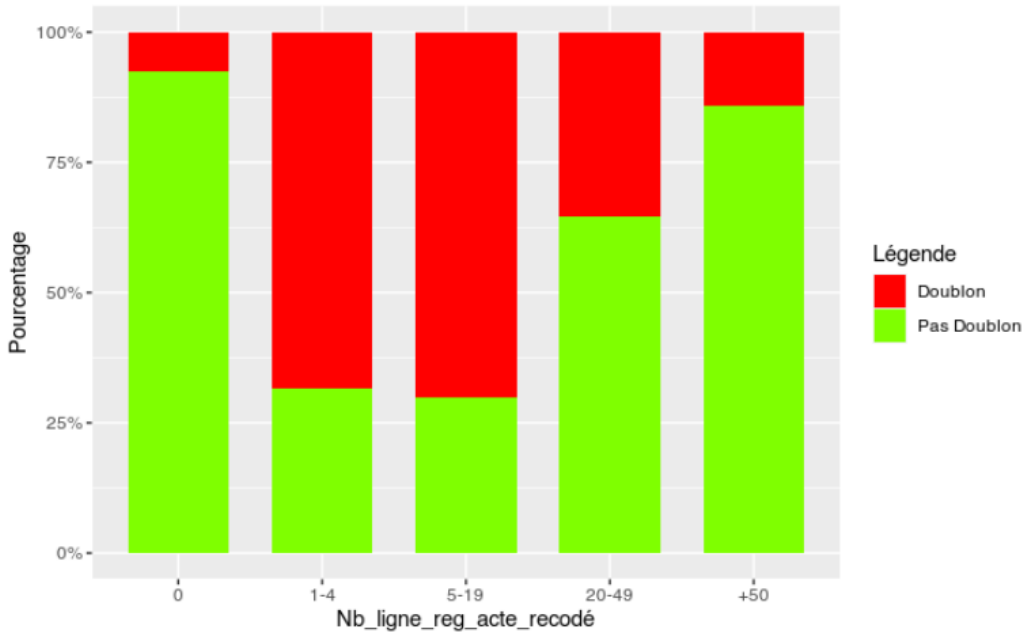


Figure 35 – Répartition de la variable Nb_ligne_reg_acte en fonction de la colonne Doublon

Les cas pour lesquels il y a des doublons et Nb_ligne _reg _acte vaut 0 correspondent en quasi-totalité à des cas de frais de séjour mal codifiés. 85% des doublons correspondent à une valeur de Nb_ligne_reg_acte comprise entre 1 et 4. Les doublons pour lesquels Nb_ligne_reg_acte est supérieur ou égal à 5 correspondent en quasi-totalité à des cas dentaires, qui donnent naturellement lieu à un plus grand nombre de lignes de soins.

La valeur des autres indicateurs créés en fonction de la variable doublon sont tracés.

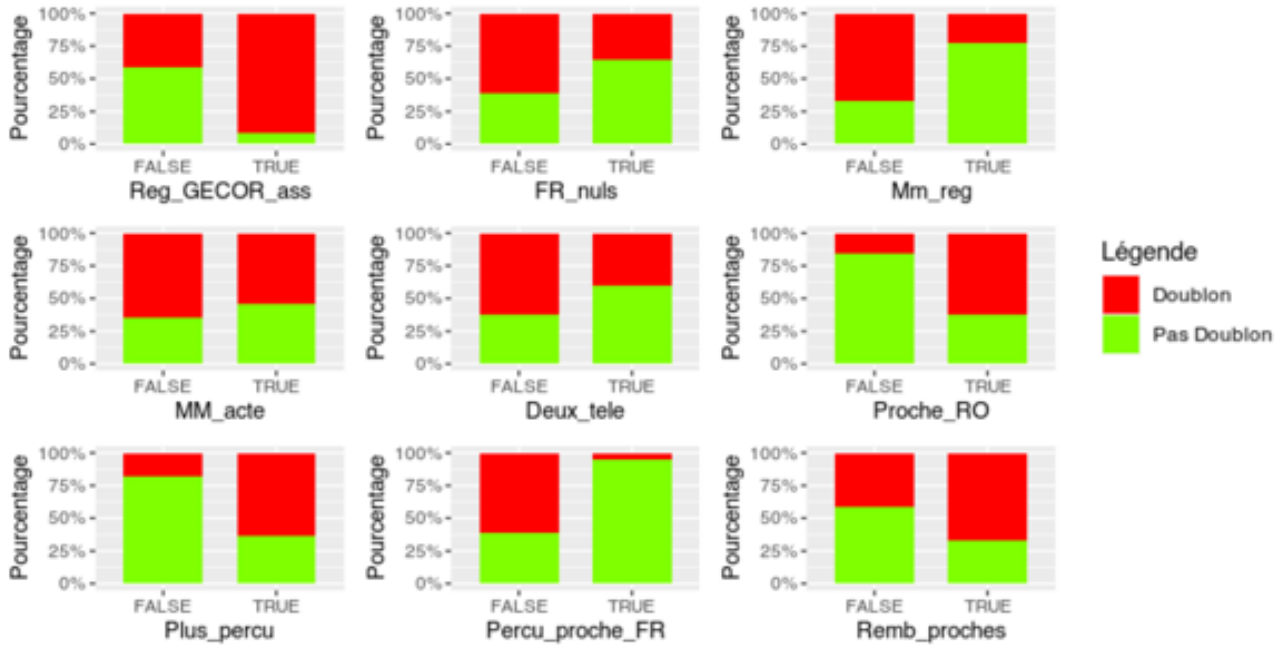


Figure 36 – Répartition des indicateurs en fonction de la colonne Doublon

A part pour la variable MM_acte, il y a un changement notable de répartition de Doublon selon la valeur de l'indicateur.

- Si un paiement est fait à l'assuré et un autre au professionnel de santé, il y a plus de chances que ce soit un doublon que si le paiement est réalisé au même type de personne.
- Si les frais réels sont nuls, il y a moins de chances d'avoir un doublon.
- Si les deux lignes de soins ont été payées via le même règlement, il y a moins de chances que la ligne concernée corresponde à un doublon.
- Si les deux lignes de soins ont été télétransmises par le régime obligatoire, il y a moins de chances qu'elles correspondent à un doublon.
- Si les remboursements du régime complémentaire sont proches, il y a plus de chances qu'il y ait un doublon. Cela s'explique par le fait que le montant du remboursement du régime complémentaire doit être identique pour deux mêmes soins.
- Si deux lignes de soins correspondent au même soin et si la somme des paiements dépasse les frais réels, cela implique une plus grande chance de doublon. En effet, si un même soin a été payé en deux temps et que l'assuré a plus perçu que les frais réels, il y a un indu.
- Si le montant perçu est proche des frais réels, il y a moins de chances de doublons. Cela indique plutôt un remboursement en deux temps.
- Si les remboursements des deux lignes de soins sont proches, il y a une plus grande chance de doublon. Cela s'explique par le fait que le montant du remboursement de la complémentaire doit être identique pour deux mêmes soins si les forfaits disponibles sont identiques.

4.5.3 Corrélations entre les variables

Les corrélations entre la variable doublon et les différents indicateurs créés sont représentées.

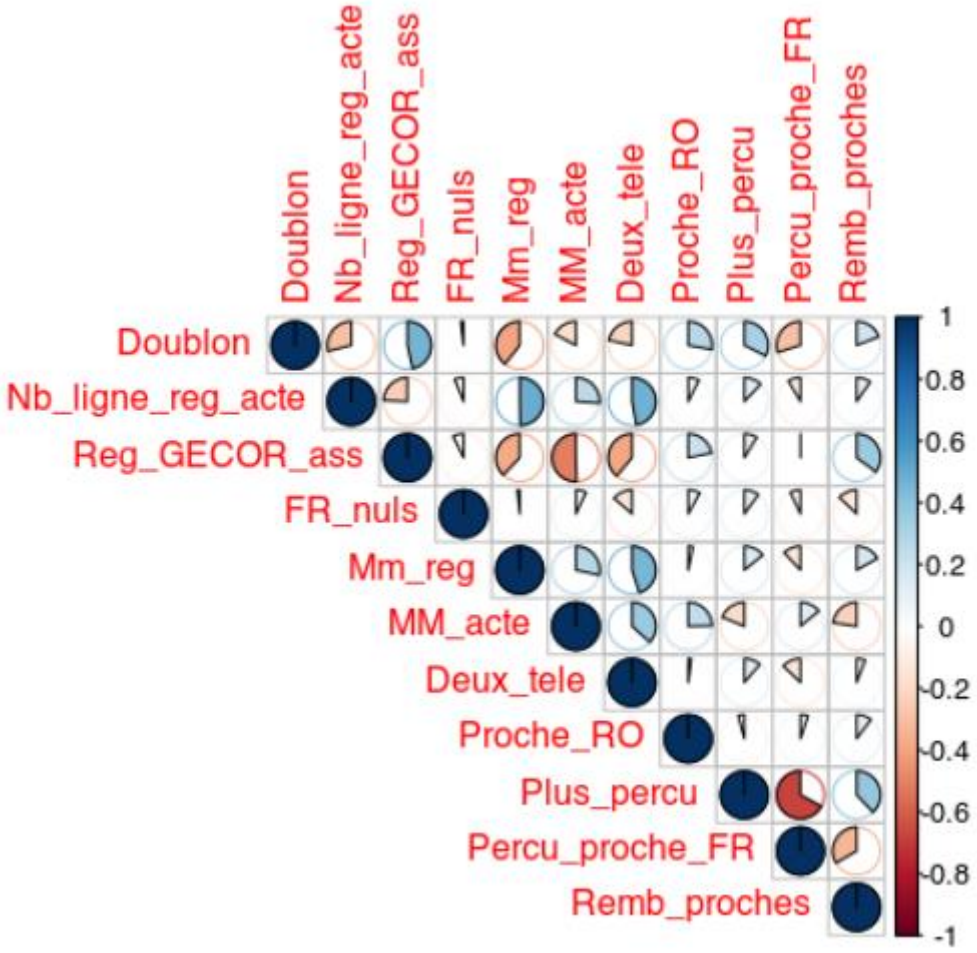


Figure 37 – Corrélations sur la base finale

Les variables Plus_percu et Percu_proche_FR sont les variables les plus corrélées. Cette corrélation s'explique pas le fait que si Plus_percu vaut 1 alors il y a plus de chances que Percu_proche_FR vaille 0 par définition des indicateurs.

Ainsi, les indicateurs représentées et la colonne indiquant le regroupement de l'acte constituent les variables explicatives.

4.6 Apprentissage supervisé et classification

Le but de cette partie est de présenter la classification supervisée, différents algorithmes et leur application à la base de données.

4.6.1 Classification supervisée et cadre de l'étude

L'apprentissage supervisé consiste en la recherche d'une règle de prédiction entre des variables explicatives et une variable expliquée.

Soit $\mathcal{X}$ le domaine des variables explicatives. Dans le cadre de ce mémoire, il y a p variables explicatives sélectionnées dans la partie précédente toutes à valeurs dans $\{0,1\}$ donc $\mathcal{X} = \{0,1\}^p$. Soit $\mathcal{Y}$ le domaine de la variable expliquée. Ici, cet ensemble est $\mathcal{Y} = \{0,1\}$.

L'apprentissage supervisé aura pour but de trouver une fonction $f: \mathcal{X} \to \mathcal{Y}$ selon un critère choisi. Pour cela, une base de données avec des valeurs observées de X et Y a été construite. Cela permettra d'appliquer la fonction f sélectionnée par le critère afin de l'appliquer à de nouvelles données X^{new} pour prédire Y^{new}.

La base de données est découpée de la manière suivante :

- 80% des données constituent la base d'apprentissage,
- 20% des données constituent la base de test.

Pour chacun des échantillons, il y a environ 60,5% de lignes correspondant à des doublons.

4.6.2 Matrice de confusion et courbe ROC

Lorsque le domaine de valeur de la variable expliquée est $\{0,1\}$, les matrices de confusion ont utilisées pour visualiser les performances d'un modèle.

Un individu est dit négatif lorsque Y=0 et positif lorsque Y=1.

		Prédiction	
		0	1
Valeur	0	Vrai négatif	Faux positif
	1	Faux négatif	Vrai positif

Figure 38 – Matrice de confusion

À partir de cette matrice, différents indicateurs sont utilisés pour évaluer le modèle.

- L'*Accuracy* : $\frac{Vrais\ négatifs+Vrais\ positifs}{Vrais\ négatifs+Vrais\ positifs+Faux\ négatifs+Faux\ positifs}$ qui correspond à la part d'éléments correctement prédis.
- Le taux de vrais positifs : $\frac{Vrais\ positifs}{Vrais\ positifs+Faux\ négatifs}$.
- Le taux de vrais négatifs : $\frac{Vrais\ négatifs}{Vrais\ négatifs+Faux\ positifs}$.
- Le taux de faux positifs : $\frac{Faux\ positifs}{Vrais\ négatifs+Faux\ positifs}$.
- La précision : $\frac{Vrais\ positifs}{Vrais\ positifs+Faux\ positifs}$.

Même si $\mathcal{Y} = \{0,1\}$, il peut être plus pertinent d'avoir $f(\mathcal{X}) \in [0,1]$. De cette manière, une probabilité d'être positif est prédite par le modèle. Dans ce cas, les modèles pourront être sélectionnés selon l'erreur quadratique moyenne.

La courbe ROC peut être utilisée dans le cas où le modèle fournit une probabilité d'appartenir à la classe positive. En fixant une valeur de référence ref, la probabilité donnée par le modèle peut être utilisée pour obtenir une prédiction dans $\{0,1\}$:

- $Y^{pred} = 1$ pour $f(x) > ref$,
- $Y^{pred} = 0$ pour $f(x) \leq ref$.

Pour tout $ref \in [0,1]$, le taux de vrais positifs et de vrais négatifs sont calculés. En traçant les points de coordonnées $(Taux\ de\ faux\ positifs, Taux\ de\ vrais\ positifs)$, la courbe ROC est obtenue. Cette courbe joint les points $(0,0)$ et $(1,1)$. Plus la courbe est concave, plus le modèle est discriminant.

L'aire sous la courbe ROC (*Area Under the Curve* AUC), est souvent utilisée pour mesurer la proximité de la courbe avec le cas d'une discrimination parfaite. Lorsque l'aire est de $\frac{1}{2}$, le modèle n'émet aucune discrimination, lorsque l'aire est 1, il y a une discrimination parfaite. Cette aire correspond à la probabilité que le modèle puisse distinguer les positifs et les négatifs.

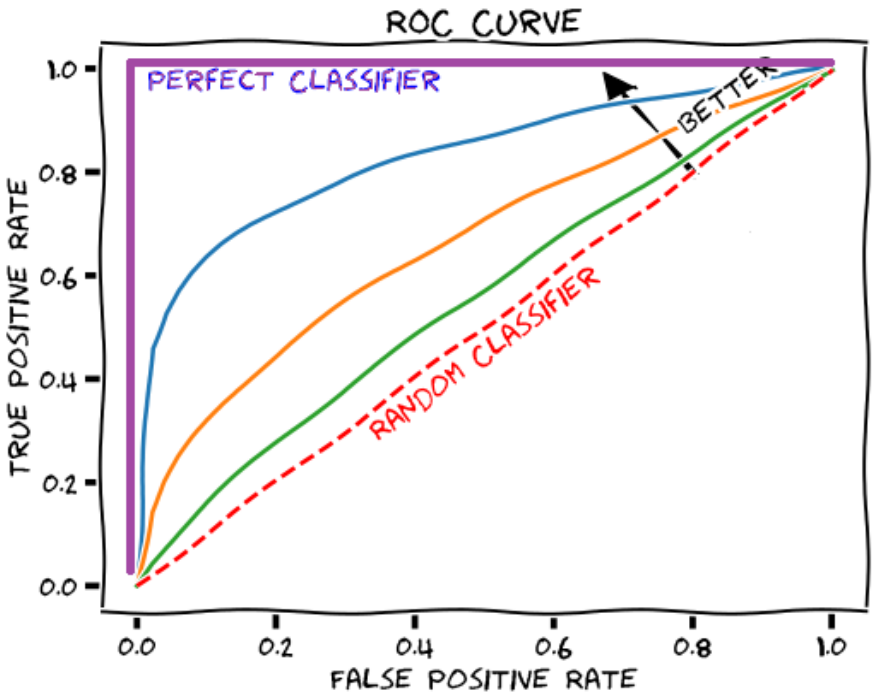


Figure 39 – Schéma explicatif du fonctionnement de la courbe ROC[8]

4.6.3 Choix des hyperparamètres

Dans chacun des algorithmes d'apprentissage supervisé utilisés, des hyperparamètres doivent être choisis pour sélectionner le modèle qui convient le mieux aux données.

Pour cela, une *grid search,* c'est-à-dire une grille qui contient plusieurs ensembles d'hyperparamètres, est utilisée. Le modèle considéré sera alors entrainé sur les différents ensembles de la grille. Dans le cadre de la classification, lorsque le modèle le permet, le critère retenu sera l'AUC. À partir des différents seuils de probabilité choisi, les matrices de confusion pourront alors être présentées.

4.6.4 Support vector machine

SVM linéaires

Description du modèle

PRINCIPE DES SVM LINEAIRES

Les *Support Vector Machines* (SVM) reposent sur la recherche de l'hyperplan optimal (lorsqu'il existe) qui permet de séparer les valeurs de $\mathcal{Y}$.

Dans ce cadre, le problème de classification est noté $\{(X_i, Y_i)_{1\leq i\leq n}\}$ avec $Y_i \in \{-1,1\}$.

Le SVM linéaire utilise un hyperplan solution d'un problème d'optimisation sous contraintes s'exprimant à l'aide de produits scalaires.

[8] Source : https://linogaliana-teaching.netlify.app/performance/

Pour cela, une fonction linéaire est utilisée :

$$f(X) = \sum_{j=1}^{p} w_j X_j + b = \langle w, X \rangle + b.$$

FONCTIONNEMENT DES SVM LINEAIRES DANS LE CAS DE DONNEES SEPARABLES

Les données d'apprentissage sont dites linéairement séparables s'il existe un hyperplan permettant de réaliser une classification parfaite. Il existe alors un nombre de séparateur linéaire infini. Pour sélectionner un séparateur linéaire, la marge du séparateur est alors introduite :

$$Marge(f) = min_i dist(X_i, H)$$

Où $dist(X_i, H)$ est la distance du point numéro i à l'hyperplan H.

Ainsi, le problème d'optimisation est le suivant :

$$f^* = argmax\ Marge(f)$$

Il s'agit alors de trouver w^*, b^*permettant de maximiser la marge.

Dans ce cadre, le problème s'écrit de la manière suivante :

$$\begin{cases} w^* = argmin \frac{1}{2}\|w\|^2 \\ \langle w, X_i \rangle + b \geq 1\ pour\ Y_i = 1 \\ \langle w, X_i \rangle + b \leq -1\ pour\ Y_i = -1 \end{cases}$$

Ce problème peut être résolu à l'aide des multiplicateurs de Lagrange.

Soit $L(w, b, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i (Y_i(\langle w, X_i \rangle + b) - 1)$.

Le problème à résoudre est alors :

$$\begin{cases} \lambda^* = argmax_\lambda (\ inf_{w,b}\ L(w, b, \lambda)\ ) \\ \lambda_i \geq 0\ pour\ tout\ i \end{cases}$$

$$\frac{\partial L}{\partial w}(w, b, \lambda) = 2 \times \frac{1}{2} \times w - \sum_{i=1}^{n} \lambda_i \big(Y_i(X_i)\big) = w - \sum_{i=1}^{n} \lambda_i Y_i X_i$$

$$\frac{\partial L}{\partial b}(w, b, \lambda) = -\sum_{i=1}^{n} \lambda_i \big(Y_i(X_i)\big) = -\sum_{i=1}^{n} \lambda_i Y_i$$

Donc si les dérivées partielles sont nulles,

$w = \sum_{i=1}^{n} \lambda_i Y_i X_i\ et \sum_{i=1}^{n} \lambda_i Y_i = 0$

Et

$$inf_{w,b}\, L(w,b,\lambda) = inf_{w,b}\, \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i (Y_i(\langle w, X_i\rangle + b) - 1)$$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i Y_i \langle \sum_{j=1}^{n} \lambda_j Y_j X_j , X_i\rangle - b\sum_{i=1}^{n} \lambda_i Y_i + \sum_{i=1}^{n} \lambda_i$$

$$= \frac{1}{2}\|w\|^2 - \langle \sum_{j=1}^{n} \lambda_j Y_j X_j , \sum_{i=1}^{n} \lambda_i Y_i X_i\rangle + \sum_{i=1}^{n} \lambda_i$$

$$= \frac{1}{2}\|w\|^2 - \|w\|^2 + \sum_{i=1}^{n} \lambda_i$$

$$= \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \langle X_i, X_j\rangle$$

Le problème s'écrit désormais :

$$\begin{cases} \lambda^* = argmax_\lambda \left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \langle X_i, X_j\rangle \right) \\ \lambda_i \geq 0 \; pour \; tout \; i \\ \sum_{i=1}^{n} \lambda_i Y_i = 0 \end{cases}$$

Ce problème peut ensuite être résolu à l'aide d'algorithmes d'optimisation quadratique.

Ce qui permet d'obtenir une solution explicite dans le cas où les données sont linéairement séparables.

FONCTIONNEMENT DES SVM LINEAIRES DANS LE CAS DE DONNEES NON SEPARABLES

Cependant, les données ne sont pas linéairement séparables en général. Une pénalité est alors introduite :

$$\varepsilon_i = \max(0, 1 - y_i(\langle w, X_i\rangle + b))$$

Le problème d'optimisation est alors le suivant :

$$\begin{cases} w^* = argmin\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i \\ \varepsilon_i \geq 0 \; pour \; tout \; i \\ Y_i(\langle w, X_i\rangle + b) \geq 1 - \varepsilon_i \; pour \; tout \; i \end{cases}$$

Où C est à choisir.

Ce problème d'optimisation peut être résolu par des méthodes identiques au cas linéairement séparable.

Dans le cas de données non séparables, le choix de C permet ainsi de prendre le modèle qui fournit les meilleurs résultats. C est choisi par le paramètre *cost* de la fonction *svm* de R. Dans le cas du SVM linéaire, le noyau est à renseigner (*kernel* = '*linear*').

SVM non linéaire

Description du modèle

FONCTIONNEMENT DES SVM NON LINEAIRES

Des surfaces séparatrices non linéaires peuvent être utilisées. Dans ce cadre, une fonction noyau est introduite. Cette fonction permet de transformer les données et de se placer dans un espace de dimension plus élevé. La justification de cette méthode se fait par le théorème de Mercer.

Théorème de Mercer :

Soit $\mathcal{X}$ un compact de R^d et K: $\mathcal{X}\times\mathcal{X}\rightarrow R$ une fonction symétrique.

Si pour tout $f \in L_2(\mathcal{X})$, $\int K(x,y)f(x)f(y)dxdy \geq 0$,

Alors il existe un espace de Hilbert H et $\Phi : \mathcal{X}\rightarrow H$ vérifiant :

$K(x,y) = \langle \Phi(x), \Phi(y)\rangle$ pour tout $x, y \in \mathcal{X}$.

$K(.,.)$ est appelée noyau défini positif.

Différents noyaux sont classiquement utilisés :

- $K(x,y) = \langle x, y\rangle$ est le noyau linéaire,
- $K(x,y) = e^{-\gamma\|x-y\|}$ est le noyau exponentiel,
- $K(x,y) = e^{-\gamma\|x-y\|^2}$ est le noyau gaussien.

Le problème d'optimisation sous contraintes s'écrit alors dans H.

$$\begin{cases} w^* = argmin\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i \\ \varepsilon_i \geq 0 \; pour \; tout \; i \\ Y_i(\langle w, \Phi(X_i)\rangle + b) \geq 1 - \varepsilon_i \; pour \; tout \; i \end{cases}$$

En utilisant les multiplicateurs de Lagrange, le problème se réécrit :

$$\begin{cases} \lambda^* = argmax_\lambda \left(-\sum_{i=1}^{n}\lambda_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j \langle \Phi(X_i), \Phi(X_j)\rangle \right) \\ C \geq \lambda_i \geq 0 \; pour \; tout \; i \\ \sum_{i=1}^{n}\lambda_i Y_i = 0 \end{cases}$$

Or, $\langle \Phi(X_i), \Phi(X_j)\rangle = K(X_i, X_j)$ donc la résolution du problème ne nécessite pas de déterminer H et Φ.

Une solution explicite peut alors être déterminée.

PARAMETRES DE L'ALGORITHME

En plus de la constante C déjà évoquée, le choix du noyau est possible par le paramètre *kernel*. Ce noyau est liée au choix du paramètre *gamma*.

Application aux données

Un SVM linéaire et un SVM non linéaire sont appliqués aux données. Pour les SVM, la sortie du modèle est 0 si la ligne ne correspond pas à un doublon et 1 si la ligne correspond à un doublon. Ainsi, les probabilités d'appartenance à la classe positive ou négative ne sont pas obtenues. La courbe ROC ne peut donc pas être utilisée. Ainsi, les autres indicateurs sont retenus pour calibrer le modèle.

Pour le SVM linéaire, le modèle de base est utilisé. Pour le SVM non linéaire, celui-ci étant plus performant sur le jeu de données, une *grid search* est utilisée pour trouver les paramètres qui conviennent le mieux.

Les résultats suivants sont obtenus :

Base d'apprentissage SVM Linéaire		Prédiction	
		0	1
Valeur	0	7373	1234
	1	1127	12097

Indicateurs base d'apprentissage	
Accuracy	89,19%
Taux de vrais positifs	91,48%
Taux de vrais négatifs	85,66%
Taux de faux positifs	14,34%
Précision	90,74%

Base de test SVM Linéaire		Prédiction	
		0	1
Valeur	0	1843	309
	1	279	3027

Indicateurs base test	
Accuracy	89,23%
Taux de vrais positifs	91,56%
Taux de vrais négatifs	85,64%
Taux de faux positifs	14,36%
Précision	90,74%

Base d'apprentissage SVM Radial		Prédiction	
		0	1
Valeur	0	8117	490
	1	572	12652

Indicateurs base d'apprentissage	
Accuracy	95,14%
Taux de vrais positifs	95,67%
Taux de vrais négatifs	94,31%
Taux de faux positifs	5,69%
Précision	96,27%

Base de test SVM Radial		Prédiction	
		0	1
Valeur	0	2034	118
	1	169	3137

Indicateurs base test	
Accuracy	94,74%
Taux de vrais positifs	94,89%
Taux de vrais négatifs	94,52%
Taux de faux positifs	5,48%
Précision	96,37%

Figure 40 – Matrices de confusion et indicateurs de performance des *Support Vector Machines*

Le SVM non linéaire correspond aux meilleures performances.

- Dans la base d'apprentissage :
  - 20769 lignes sont bien classées. Ces lignes correspondent à 12652 doublons et 8117 non doublons.
  - 572 doublons sont classés comme non doublons et 490 non doublons sont classés comme doublons.
  - L'*accuracy* est 95,14%.
- Dans la base de test :
  - 5171 lignes sont bien classées. Ces lignes correspondent à 3137 doublons et 2034 non doublons.
  - 169 doublons sont classés comme non doublons et 118 non doublons sont classés comme doublons.
  - L'*accuracy* est 94,74%.

Ce modèle présente de bonnes performances. Cependant, des modèles permettant d'avoir une probabilité en sortie du modèle seront choisis. En effet, en cas de contrôles de doubles règlements, il se peut que toutes les lignes ne soient pas analysées par manque de temps. Ainsi, un modèle ayant en sortie des probabilités pourra permettre de prendre les probabilités les plus élevées et donc d'analyser les lignes qui ont le plus de chances d'être des doublons.

4.6.5 Forêts aléatoires

Arbres CART

Description du modèle

PRINCIPE DE L'ALGORITHME

Les arbres de décision permettent de classer des individus à l'aide de séquences de tests. Un arbre CART repose sur des tests binaires et fonctionne pour des variables quantitatives.

Dans ce cadre, différents éléments sont introduits :

- La racine correspond à l'ensemble de la population, à l'échantillon initial non encore séparé,
- Les branches correspondent aux règles qui séparent en deux la population,
- Les nœuds correspondent aux sous-échantillons créés,
- Les feuilles correspondent aux sous-population homogènes donnant une estimation.

FONCTIONNEMENT DE L'ALGORITHME

Pour l'algorithme CART, un arbre est construit par séparation itérative des nœuds. Chaque nœud est séparé en deux feuilles si ce dernier n'est pas assez pur. Le but est d'obtenir deux nœuds filles les plus purs possibles. Dans le cadre de la classification, chaque feuille obtient la valeur de la classe majoritaire. Un indice d'impureté et un critère de séparation sont alors définis.

Pour la classification binaire, l'indice d'impureté de GINI peut être utilisé. Pour un nœud de valeur S, $I_{GINI}(S) = p_0(1-p_0) + p_1(1-p_1)$.

Avec :

- p_0 la proportion de Y prenant la valeur 0 dans S,
- p_1 la proportion de Y prenant la valeur 1 dans S.

Plus une classe est majoritaire, plus cet indice est petit.

L'indice d'impureté de l'entropie peut également être utilisé en classification binaire. Pour un nœud de valeur S, $I_{Entropie}(S) = -p_0 \log(p_0) - p_1 \log(p_1)$. De même, plus une classe est majoritaire, plus cet indice est petit.

En régression, l'indice d'impureté de la variance est utilisé.

Pour un nœud de valeur S, $I_{Variance}(S) = var(Y(S))$.

L'algorithme va parcourir l'ensemble des possibles et choisir le couple $(variable, seuil)$ séparant S en S_1, S_2 de tailles n_1, n_2 qui maximisent :

- $I(S) - \frac{n_1}{n_1+n_2} I(S_1) - \frac{n_2}{n_1+n_2} I(S_2)$ en classification,
- $var\big(Y(S)\big) - var\big(Y(S_1)\big) - var(Y(S_2))$ en régression.

Cette séparation est effectuée sur chaque nœud jusqu'à ce que $I(S) = 0$ pour tout S (élagage).

L'arbre CART possède plusieurs inconvénients. Même s'il est très lisible et permet une bonne interprétation, il est instable en cas de nombreuses variables ou de variables corrélés.

Forêts aléatoires

Description du modèle

Principe de l'algorithme

Pour pallier les problèmes des arbres, les forêts aléatoires ont été introduites.

Une forêt est constituée d'un ensemble d'arbre effectuant une prédiction.

Les forêts aléatoires reposent sur une amélioration du *bagging* (abréviation de *boostrap aggregating*) en utilisant les arbres CART.

Fonctionnement du *bagging*

Le *bagging* utilise la prévision de modèles indépendants dans le but de réduire l'erreur de prédiction. Il a été introduit par L. Breiman en 1996 Pour réduire la dépendance entre les modèles, le *boostrap* est utilisé.

N réplications *bootstrap* de l'échantillon sont obtenus par des tirages avec remise. Sur chacun de ses échantillons, une estimation $\hat{f}_i$ est obtenue.

La prévision est enfin obtenue de la manière suivante :

- $\hat{f} = \frac{1}{N}\sum_{i=1}^{N} \hat{f}_i$ pour une variable quantitative,
- $\hat{f} = argmax_K\big(card\{\, i \mid \hat{f}_i = K\}\big)$ pour une variable qualitative.

La prédiction finalement sélectionnée correspond ainsi à la classe majoritaire en classification ou à la moyenne des prédictions en régression.

FONCTIONNEMENT DES FORETS ALEATOIRES

Breiman introduit en 2001 une amélioration du *bagging* pour les arbres CART.

Les forêts aléatoires améliorent le *bagging* en rendant les arbres plus indépendants. Les variables sont sous-échantillonnées. A chaque nœud, un échantillon aléatoire des variables est considéré. Cela permet de réduire le sur-apprentissage.

N réplications *bootstrap* de l'échantillon sont obtenus par des tirages avec remise.

Sur chacun de ces échantillons, l'algorithme CART est appliqué avec une variante : à chaque fois qu'un nœud est séparé en deux, un sous-échantillonnage des variables est effectué et le meilleur découpage est sélectionné uniquement avec les variables sélectionnées.

La prévision est ainsi obtenue par la même formule que le *bagging*.

L'ajout du sous-échantillonnage permet d'avoir des arbres moins corrélés. C'est aussi pour cela qu'en général l'algorithme utilise des arbres peu profonds.

ERREUR *OUT-OF-BAG*

L'erreur *out-of-bag* repose sur le fait que pour chaque arbre, une partie des observations n'est pas utilisée.

Pour chaque observation (X_i, Y_i), une prédiction est calculée avec les arbres pour lesquels l'échantillon *boostrap* qui a servi à la construction ne contient pas X_i :

$$\hat{Y}_i = \frac{1}{card(A_i)} \sum_{j \in A_i} \hat{f}_k(X_i)$$

Avec A_i l'ensemble des arbres de la forêt aléatoire pour lesquels l'échantillon *boostrap* qui a servi à la construction ne contient pas X_i.

Le calcul de l'erreur out-of-bag s'effectue alors de la manière suivante :

- $E_{OOB} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ dans le cas de la régression,
- $E_{OOB} = \frac{1}{n}\ card\left\{i \mid \hat{Y}_i \neq Y_i\right\}$ dans le cas de la classification.

IMPORTANCE DES VARIABLES

Breiman & Cutler proposent en 2005 une méthode pour déterminer les variables les plus importantes dans le modèle.

Soient $X^{(1)}, \ldots, X^{(p)}$ les variables explicatives du modèle.

Soit OOB_k l'échantillon des observations qui n'appartiennent pas à l'échantillon *bootstrap* utilisé pour construire l'arbre CART k.

L'erreur de prédiction de l'arbre est définie par :

$\mathrm{E}_{\mathrm{OOB_k}} = \frac{1}{\mathrm{card(OOB_k)}} \sum_{\mathrm{i} \in \mathrm{OOB_k}} (\hat{f}_k(X_i) - Y_i)^2$.

Soit OOB_k^j l'échantillon OOB_k pour lequel les valeurs prises par la variable j ont été permutées de façon aléatoire et X_i^j ces observations après permutation.

L'erreur de prédiction après permutation sur cet arbre est définie par :

$\mathrm{E}^{\mathrm{j}}_{\mathrm{OOB_k}} = \frac{1}{\mathrm{card(OOB_k)}} \sum_{\mathrm{i} \in \mathrm{OOB_k}} (\widehat{f_k}(X_i^j) - Y_i)^2.$

L'importance de la variable $X^{(j)}$ est alors donnée par :

$$\mathrm{Imp}_{X^{(j)}} = \frac{1}{\mathrm{N}} \sum_{k=1}^{N} (\mathrm{E}_{\mathrm{OOB_k}} - \mathrm{E}^{\mathrm{j}}_{\mathrm{OOB_k}}).$$

Plus $\mathrm{Imp}_{X^{(j)}}$ est grande, plus la variable $X^{(j)}$ est importante. En effet, plus la variable $X^{(j)}$ est importante, plus une permutation de ses valeurs impactera l'erreur.

PARAMETRES DE L'ALGORITHME

Le package *caret* propose plusieurs paramètres pour construire la forêt aléatoire.

- Le paramètre *mtry* est le nombre de variable candidates dans les sous-échantillons créés à chaque nœud.
- *Min.node.size* correspond au nombre minimum d'observations dans les feuilles de chaque arbre.
- Le paramètre *splitrule* est l'indice d'impureté utilisé pour la construction des arbres.

Application aux données

Les forêts aléatoires permettent d'obtenir des probabilités en sortie du modèle. Ainsi, la courbe ROC peut être tracée et sera un critère important pour calibrer le modèle en faisant le choix parmi les hyperparamètres.

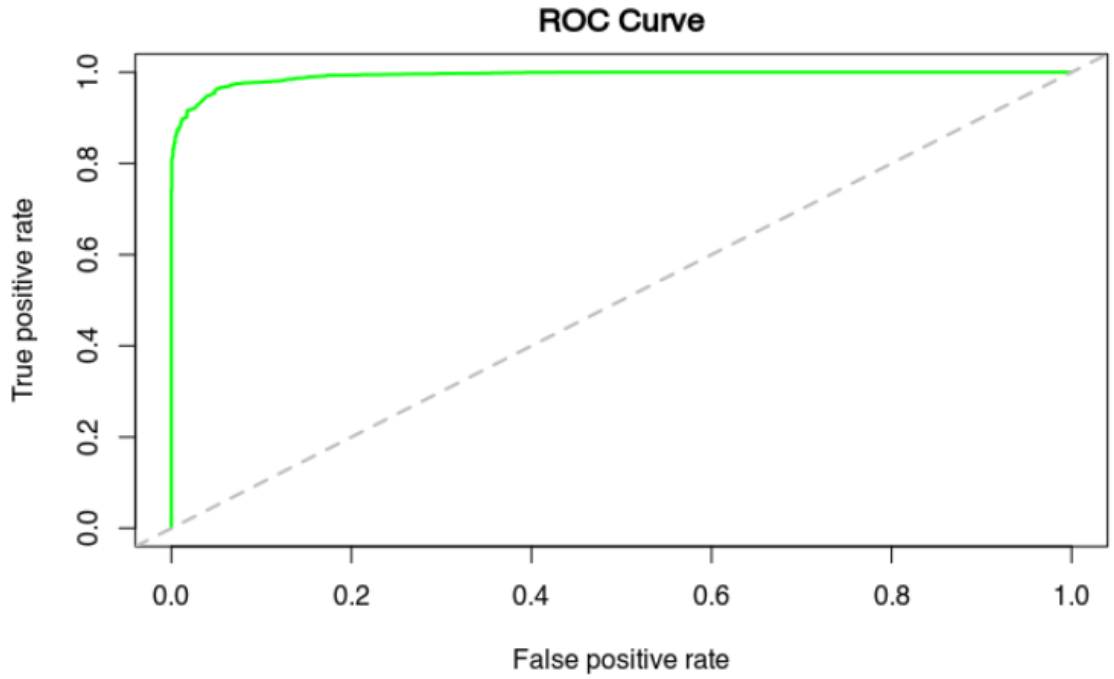


Figure 41 – Courbe ROC du modèle *Random Forest* sur les données test

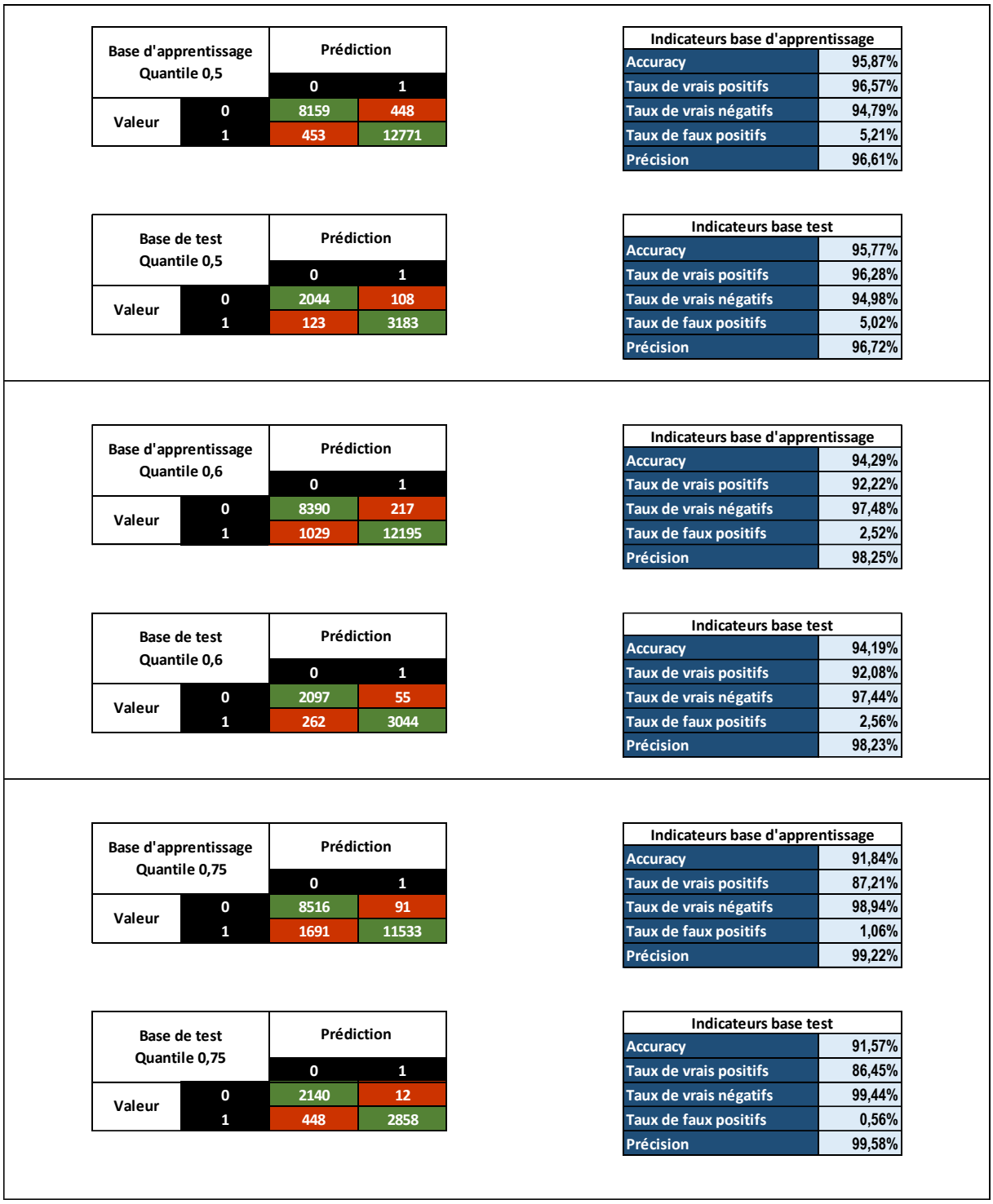
L'aire sous la courbe obtenue pour le meilleur modèle est 0,9923 pour le jeu de données test ce qui indique une très bonne performance.

Avec les probabilités, un seuil doit être choisi pour définir si la ligne correspond à un doublon ou non.

En notant s le seuil choisi :

- Si $prob_{modèle} > s$, la ligne est considérée comme doublon,
- Sinon la ligne est considérée comme non doublon.

Après analyse des résultats données par différents seuils, trois seuils sont sélectionnés : $s \in \{0{,}5\ ;\ 0{,}6\ ;\ 0{,}75\}$. Pour ces trois seuils, les matrices de confusion et les indicateurs de performances sont présentés.



Base d'apprentissage Quantile 0,5		Prédiction	
		0	1
Valeur	0	8159	448
	1	453	12771

Indicateurs base d'apprentissage	
Accuracy	95,87%
Taux de vrais positifs	96,57%
Taux de vrais négatifs	94,79%
Taux de faux positifs	5,21%
Précision	96,61%

Base de test Quantile 0,5		Prédiction	
		0	1
Valeur	0	2044	108
	1	123	3183

Indicateurs base test	
Accuracy	95,77%
Taux de vrais positifs	96,28%
Taux de vrais négatifs	94,98%
Taux de faux positifs	5,02%
Précision	96,72%

Base d'apprentissage Quantile 0,6		Prédiction	
		0	1
Valeur	0	8390	217
	1	1029	12195

Indicateurs base d'apprentissage	
Accuracy	94,29%
Taux de vrais positifs	92,22%
Taux de vrais négatifs	97,48%
Taux de faux positifs	2,52%
Précision	98,25%

Base de test Quantile 0,6		Prédiction	
		0	1
Valeur	0	2097	55
	1	262	3044

Indicateurs base test	
Accuracy	94,19%
Taux de vrais positifs	92,08%
Taux de vrais négatifs	97,44%
Taux de faux positifs	2,56%
Précision	98,23%

Base d'apprentissage Quantile 0,75		Prédiction	
		0	1
Valeur	0	8516	91
	1	1691	11533

Indicateurs base d'apprentissage	
Accuracy	91,84%
Taux de vrais positifs	87,21%
Taux de vrais négatifs	98,94%
Taux de faux positifs	1,06%
Précision	99,22%

Base de test Quantile 0,75		Prédiction	
		0	1
Valeur	0	2140	12
	1	448	2858

Indicateurs base test	
Accuracy	91,57%
Taux de vrais positifs	86,45%
Taux de vrais négatifs	99,44%
Taux de faux positifs	0,56%
Précision	99,58%

Figure 42 – Matrices de confusion et indicateurs de performance du *Random Forest*

Plus le seuil augmente, plus la précision augmente. C'est le résultat espéré. En effet, seuls les cas prédits comme doublons par le modèle seront analysés pour traitement. Si un nombre réduit de cas doivent être analysés, les lignes qui correspondent à une plus grande probabilité d'être un doublon sont choisies. Augmenter le seuil conduit à diminuer le pourcentage de faux positifs, ce qui convient. Néanmoins, cela ne permet pas d'augmenter l'*accuracy*, ou le taux de vrais positifs.

Les importances des variables sont représentées :

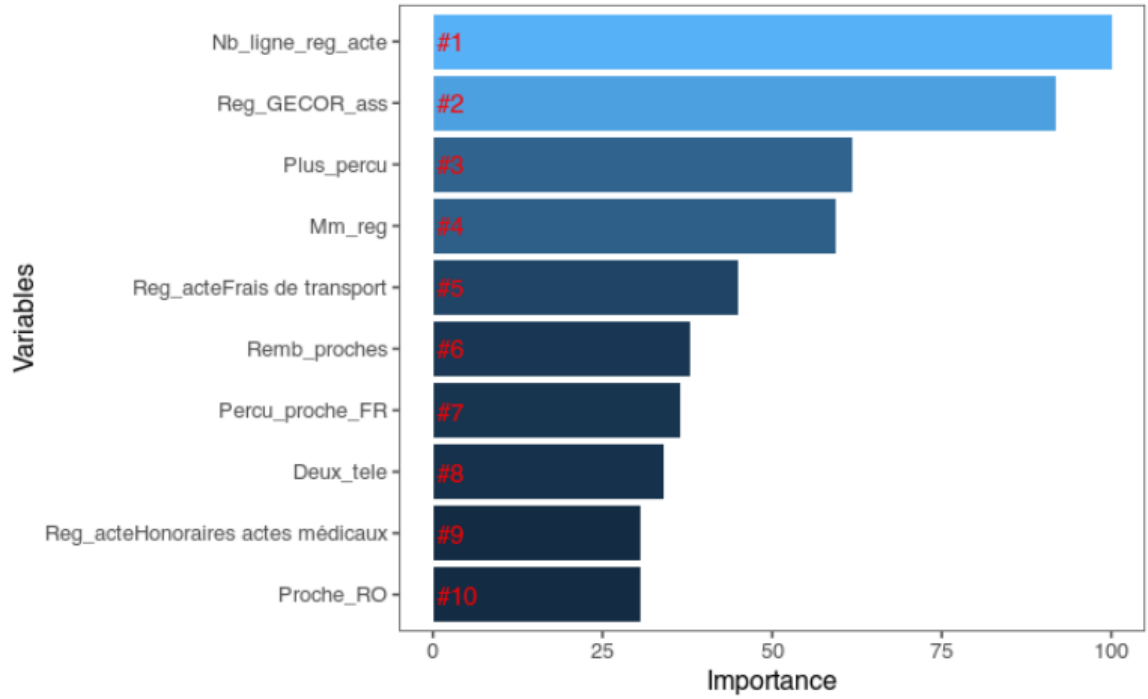


Figure 43 – Importance des variables du modèle *Random Forest*[9]

Les variables les plus importantes du modèle sont :

- La variable Nb_ligne_reg_acte,
- La variable Reg_GECOR_ass
- La variable Plus_percu.

Ces variables correspondent à ce qui était attendu.

Dans la suite, seuls les résultats pour le seuil $s = 0{,}5$ seront présentés, ce seuil étant celui qui a été sélectionné au final.

4.6.6 Modèles reposant sur le *Gradient boosting*

Gradient boosting

Description du modèle

PRINCIPE DU *GRADIENT BOOSTING*

Le *gradient boosting* a été introduit par Friedman en 2002.

[9] Basé sur https://www.kaggle.com/code/arunkumarramanan/data-science-in-r-and-titanic-survival-prediction

Le *gradient boosting* repose sur une séquence de modèles peu performants qui à chaque étape permettent d'aller vers une meilleure solution. Chaque modèle est entraîné de manière à corriger les erreurs du modèle précédent.

Cet algorithme utilise la descente de gradient.

ALGORITHME DE DESCENTE DE GRADIENT

L'algorithme de descente de gradient a pour but de calculer le minimum d'une fonction.

Soient $f : R^n \to R$ différentiable et $\varepsilon > 0$ le niveau d'erreur.

L'algorithme de descente de gradient est itératif :

$$\begin{cases} P_0 = (x_1, \ \dots, \ x_n) \text{ le point initial} \\ P_{k+1} = P_k - \delta_k \times \nabla f(P_k) \end{cases}$$

Où à chaque itération, le pas δ_k est choisi.

L'algorithme s'arrête lorsque $\|\nabla f(P_k)\| < \varepsilon$.

Pour le choix du pas, l'idéal serait d'avoir $f(P_{k+1}) \leq f(P_k)$.

Le pas peut être constant, dans ce cas $\delta = \delta_k$ est appelé *learning rate*.

L'algorithme peut être appliqué pour minimiser des fonctions de perte.

FONCTIONNEMENT DU *GRADIENT BOOSTING*

Le *gradient boosting* constitue un ensemble de modèles de la manière suivante :

$$\widehat{f_k} = \widehat{f_{k-1}} - \gamma_k \sum_{i=1}^{n} \nabla l\left(Y_i, \widehat{f_{k-1}}(X_i)\right)$$

Où $l(.,.)$ est une fonction de perte.

Le modèle est initialisé de la manière suivante :

$$\widehat{f_0} = argmin_\gamma \sum_{i=1}^{n} l(Y_i, \gamma)$$

Soit N le nombre d'arbres utilisés.

Pour $k \in [\![1, N]\!]$:

- Les résidus sont calculés par la formule $r_{k,i} = -\left[\frac{\partial l(Y_i, f(X_i))}{\partial f(X_i)}\right]_{f=\widehat{f_{k-1}}}$,
- Un arbre a_k est utilisé pour la prédiction des couples $\left(X_i, r_{k,i}\right)_{i \in [\![1,n]\!]}$,
- La fonction est mise à jour par $\widehat{f_k} = \widehat{f_{k-1}} - \gamma_k a_k$ où $\gamma_k = argmin_\gamma \sum_{i=1}^{n} l\left(Y_i, \widehat{f_{k-1}}(X_i) - \gamma a_k(X_i)\right)$.

La prédiction est enfin obtenue par $\widehat{f_N}(X)$.

Pour diminuer le sur-ajustement, un paramètre *shrinkage* peut être utilisée. Dans ce cas, $\widehat{f_k} = \widehat{f_{k-1}} - \eta\, \gamma_k a_k$ avec $\eta \in]0,1[$. Plus ce paramètre est petit, plus le nombre d'arbres augmente.

Pour diminuer le sur-ajustement, une autre possibilité proposée par Friedman (2002) est d'utiliser un sous-échantillonnage à chaque étape pour construire des prédicteurs plus indépendants. C'est le *stochastic gradient boosting*.

PARAMETRES DE L'ALGORITHME

Le package caret effectuant le *gradient boosting* propose différents paramètres qu'il convient d'optimiser.

- Le paramètre distribution correspond à la distribution utilisée. Les distributions suivantes peuvent par exemple être utilisées :
  - *Gaussian* qui correspond au carré des erreurs,
  - *Bernoulli* pour la fonction de coût logistique,
  - Laplace pour la valeur absolue
- *N.trees* correspond au nombre d'arbres utilisés.
- Les arbres ont une profondeur donnée par le paramètre *interaction.depth*.
- Le nombre minimum d'observations dans les feuilles de chaque arbre est donnée par *n.minobsinnode*.
- Le paramètre *shrinkage* déjà évoqué peut être modifié.

Application aux données

Les modèles de *gradient boosting* sont plus coûteux en termes de calculs. L'apprentissage est plus long à effectuer que pour les forêts aléatoires.

Pour le modèle avec les meilleurs paramètres, l'aire sous la courbe ROC obtenue avec les données test est 0,9915.

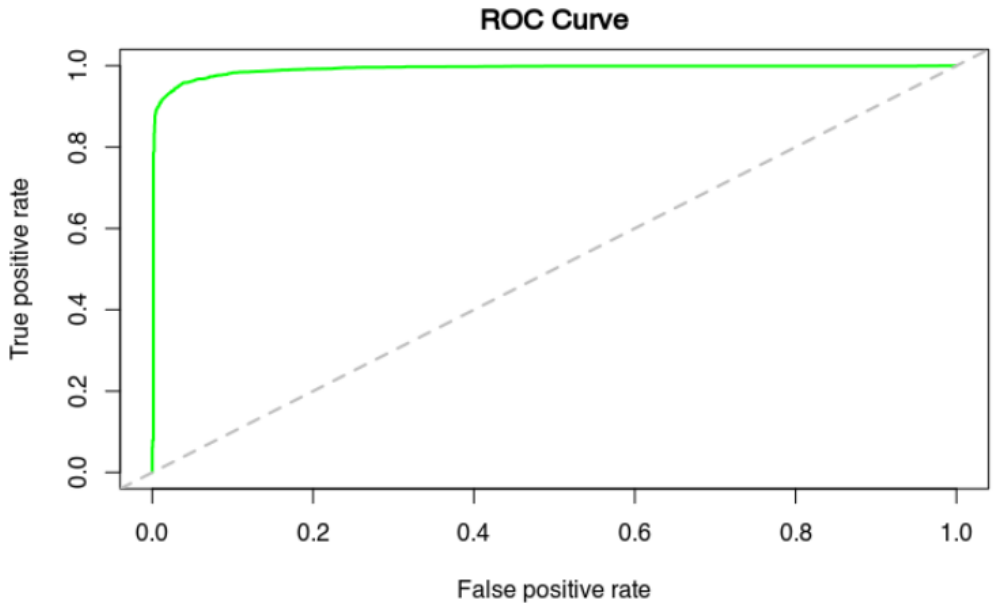


Figure 44 – Courbe ROC du modèle *Gradient Boosting* sur les données test

Le *gradient boosting* fournit également de bons résultats.

Les résultats sont très proches de ceux obtenus par *Random Forest*. L'accuracy quasi identique. Cependant, la précision est plus faible ce qui conduirait à sélectionner le modèle *Random Forest*.

Base d'apprentissage Quantile 0,5		Prédiction 0	Prédiction 1
Valeur	0	8118	489
	1	409	12815

Indicateurs base d'apprentissage	
Accuracy	95,89%
Taux de vrais positifs	96,91%
Taux de vrais négatifs	94,32%
Taux de faux positifs	5,68%
Précision	96,32%

Base de test Quantile 0,5		Prédiction 0	Prédiction 1
Valeur	0	2034	118
	1	113	3193

Indicateurs base test	
Accuracy	95,77%
Taux de vrais positifs	96,58%
Taux de vrais négatifs	94,52%
Taux de faux positifs	5,48%
Précision	96,44%

Figure 45 – Matrices de confusion et indicateurs de performance du *Gradient Boosting*

Depuis son invention, le *gradient boosting* a donné place à différentes variantes.

Le *XGBOOST* est présenté dans la suite de la partie.

XGBOOST

Description du modèle

PRINCIPE DE L'ALGORITHME

L'algorithme *XGBOOST* repose sur un principe similaire au *gradient boosting*.

Il a été introduit par Chen et Guestrin en 2016.

Une nouvelle fonction objectif comprenant un terme de régularisation est utilisée.

A l'étape k, cette fonction a la forme suivante :

$$OBJ_k = \sum_{i=1}^{n} l\left(Y_i, \widehat{f}_k(X_i)\right) + \sum_{j=1}^{t} \Omega(\widehat{f}_j).$$

Le terme de régularisation permet de limiter le sur ajustement.

PARAMETRES DE L'ALGORITHME

L'algorithme XGBOOST contient un grand nombre de paramètres.

Dans le cadre du mémoire, le package *caret* est utilisé, les paramètres suivants sont utilisés pour la sélection du modèle :

- *Nrounds* est le nombre d'arbres utilisés.
- Les arbres ont une profondeur donnée par le paramètre *max_depth*.
- *Eta* est le *learning rate*.
- *Gamma* est un paramètre utilisé pour le terme de régularisation dans la fonction objectif.
- Le paramètre *colsample_bytree* est le nombre de variable candidates dans les sous-échantillons créés à chaque nœud.
- Le nombre minimum d'observations dans les feuilles de chaque arbre est donnée par *min_child_weight*.
- Le paramètre subsample correspond au pourcentage de variables tirés aléatoirement à chaque étape pour la construction des arbres.

Application aux données

Après sélection des hyperparamètres avec une *grid search*, *XGBOOST* fournit de meilleures performances. L'AUC sur le jeu de données test est 0,9938.

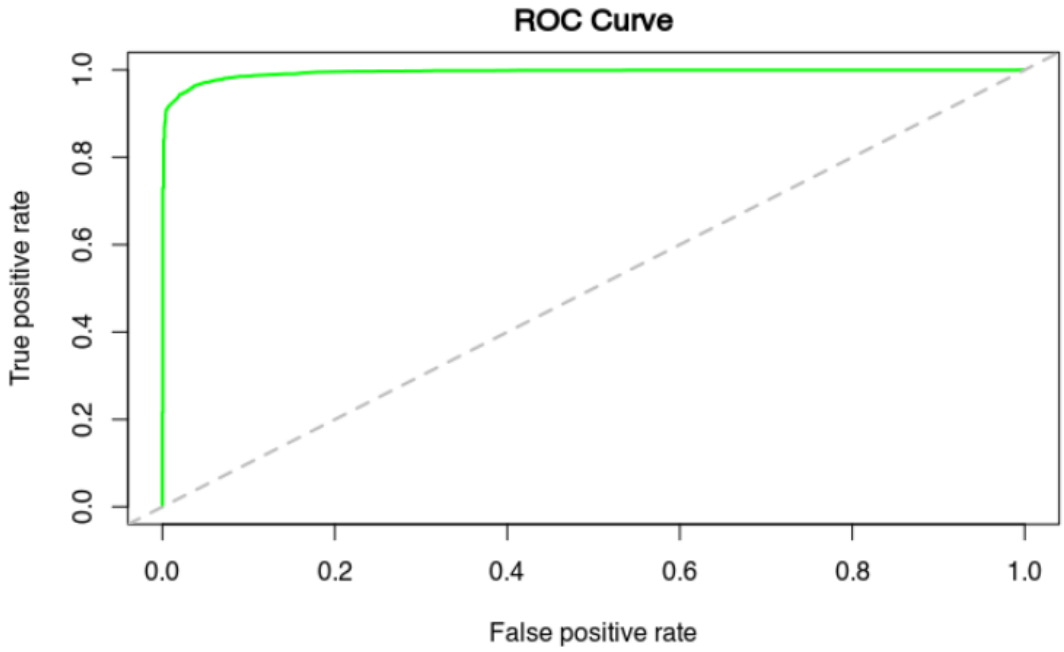


Figure 46 – Courbe ROC du modèle *XGBOOST* sur les données test

Base d'apprentissage Quantile 0,5		Prédiction	
		0	1
Valeur	0	8200	407
	1	351	12873

Indicateurs base d'apprentissage	
Accuracy	96,53%
Taux de vrais positifs	97,35%
Taux de vrais négatifs	95,27%
Taux de faux positifs	4,73%
Précision	96,94%

Base de test Quantile 0,5		Prédiction	
		0	1
Valeur	0	2055	97
	1	103	3203

Indicateurs base test	
Accuracy	96,34%
Taux de vrais positifs	96,88%
Taux de vrais négatifs	95,49%
Taux de faux positifs	4,51%
Précision	97,06%

Figure 47 – Matrices de confusion et indicateurs de performance du *XGBOOST*

L'*accuracy* et la précision sont plus élevés pour le *XGBOOST* que le *Gradient Boosting* et *Random Forest*.

En prenant en compte le temps de calcul et d'optimisation des différents paramètres, *Random Forest* pourrait être préférable. En effet, *XGBOOST* est plus coûteux en temps de calculs.

4.6.7 Les réseaux de neurones

Description du modèle

PRINCIPE DE L'ALGORITHME

Un réseau de neurones est constitué d'un ensemble de neurones liés les uns aux autres. Ils ont été introduits par Warren McCullough et le Walter Pitts en 1943. Le réseau de neurones s'inspire de la manière dont fonctionne le cerveau humain.

NEURONE BIOLOGIQUE ET NEURONE ARTIFICIEL

Un neurone biologique est une cellule du système nerveux.

Il contient :

- Un axone transmettant des messages à l'aide de signaux électriques,
- Des synapses qui lient plusieurs neurones,
- Des dendrites qui reçoivent l'influx nerveux transmis par les synapses.

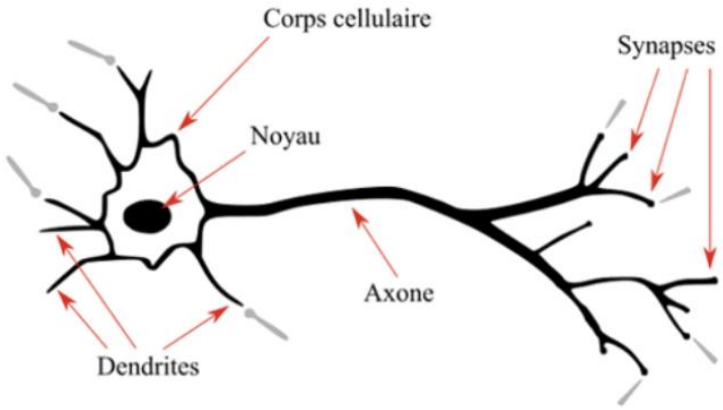


Figure 48 – Représentation d'un neurone biologique[10]

Le neurone artificiel a pour but de mimer ce processus.

[10] Source : https://deeplylearning.fr/cours-theoriques-deep-learning/fonctionnement-du-neurone-artificiel/

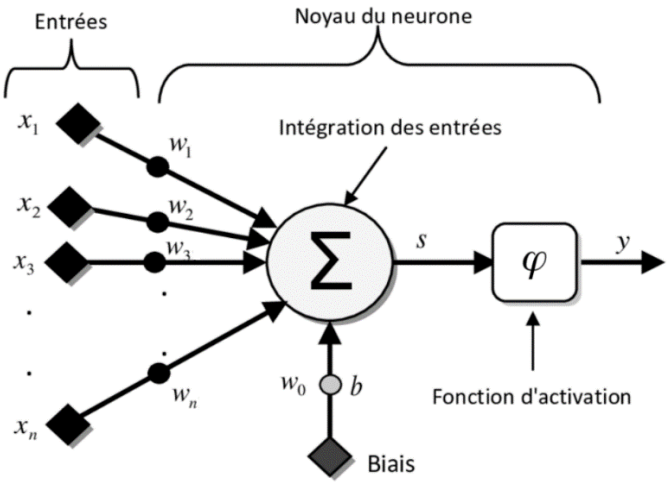


Figure 49 – Représentation d'un neurone artificiel[11]

Les notations suivantes seront utilisées :

- $(x_1, \dots, x_n)$ les valeurs d'entrée,
- Le biais b,
- $(w_0, \dots, w_n)$ les poids synaptiques,
- φ la fonction d'activation,
- y la sortie.

Les signaux des neurones sont agrégés : $s = w_0 b + \sum_{i=1}^{n} w_i x_i$.

Ensuite, la fonction d'activation est appliquée : $y = \varphi(s)$.

Différentes fonctions d'activations peuvent être utilisées. Le tableau ci-dessous regroupe les principales fonctions d'activation.

Nom de la fonction d'activation	Formule
Identité	x
Fonction à seuil	$1_{\{x \geq 0\}}$
Logistique	$\frac{1}{1 + \exp(-x)}$
Rectified linear unit	$max(0, x)$

Figure 50 – Exemples de fonctions d'activation

Dans le cadre du mémoire, les réseaux de neurones *feedforward* sont considérés. Ces réseaux lient plusieurs neurones entre elles pour lesquels la propagation se fait uniquement vers l'avant.

[11] Source : https://www.researchgate.net/figure/Modele-dun-neurone-artificiel_fig30_324929383

LES PERCEPTRONS MULTICOUCHES

Dans un réseau de neurones de types perceptron multicouche, l'information circule à travers plusieurs couches de neurones uniquement vers l'avant.

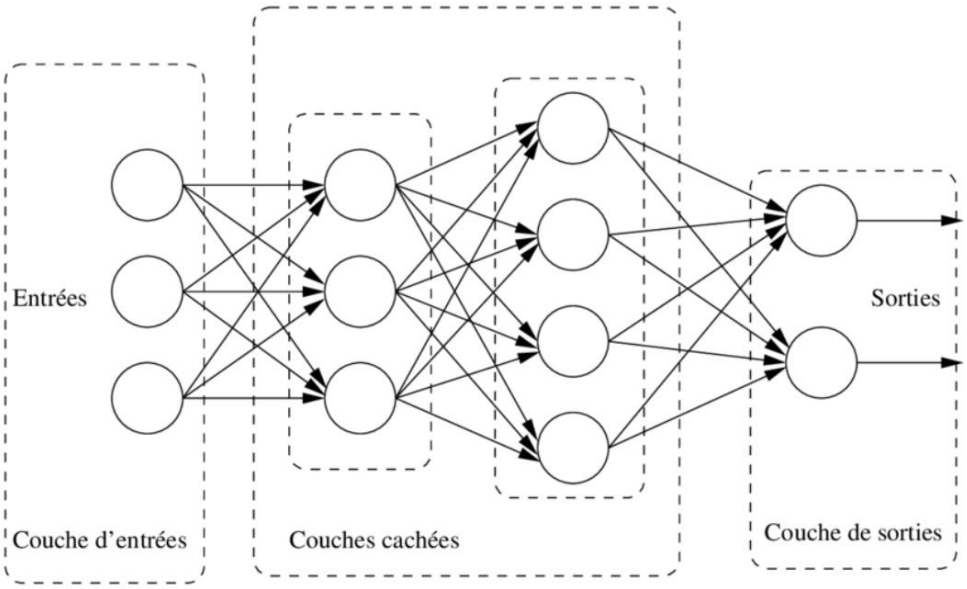


Figure 51 – Représentation d'un réseau de neurone perceptron multicouche[12]

Comme il est possible de voir sur ce schéma, le réseau de neurones se compose de trois types de couches.

- La couche d'entrées correspond aux données d'entrée du réseau de neurones.
- Les couches cachées correspondent aux couches de neurones internes au réseau.
- La couche de sortie a pour rôle de synthétiser l'information en donnant la sortie du réseau.

Soient :

- K le nombre de couches du réseau de neurones,
- $w_{i,j}^{(k)}$ le poids entre le neurone i de la couche $k-1$ et le neurone j de la couche k,
- n_k le nombre de neurones de la couche k,
- φ_k la fonction d'activation commune aux neurones de la couche k,
- $s_j^{(k)}$ la fonction d'agrégation du neurone j et la couche k,
- $x_j^{(k)}$ la valeur du neurone j de la couche k.

[12] Source : https://tel.archives-ouvertes.fr/tel-00260013/document

Le signal est propagé en avant dans les couches du réseau selon les étapes suivantes :

- $x^{(0)}$ correspond aux valeurs d'entrée.
- Pour k allant de 1 à K,

$$x_j^{(k)} = \varphi_k\left(s_j^{(k)}\right) = \varphi_k\left(\sum_{i=1}^{n_{k-1}} w_{i,j}^{(k)} x_i^{(k-1)}\right).$$

- À la fin de la propagation, la sortie $\hat{y}$ est obtenue.

RETROPROPAGATION DES ERREURS

Dans le cadre du perceptron multicouche, l'algorithme de rétropropagation des erreurs est utilisé pour calibrer les poids de manière itérative. Cette méthode utilise l'algorithme de descente du gradient.

L'initialisation de l'algorithme se fait avec des poids w aléatoires.

Il y a ensuite propagation vers l'avant selon les étapes décrites et la sortie $\hat{y}$ est obtenue.

L'erreur entre la sortie et y est calculée :

$$e_i^{sortie} = \varphi'_{sortie}\left(s_i^{sortie}\right)(y_i - \hat{y}_i).$$

Puis l'erreur est propagée vers l'arrière :

$$e_i^{(k-1)} = \varphi'_{k-1}\left(s_i^{(n-1)}\right)\sum_{j=1}^{n_k} w_{i,j}^{(k)} e_j^{(k)}.$$

Et les poids sont mis à jour :

$$w_{i,j}^{(l)} = w_{i,j}^{(l)} - \lambda e_i^{(l)} x_i^{(l-1)}.$$

Avec λ le *learning rate*.

PARAMETRES DE L'ALGORITHME

Le package *caret* est utilisé avec la méthode *nnet*.

- *Size* correspond au nombre de neurones dans la couche cachée,
- *Decay* est un paramètre de régularisation intervenant dans la fonction coût qui permet de réduire le sur-apprentissage,
- *Maxit* correspond au nombre maximum d'itérations,
- *Linout* permet de choisir entre les fonctions d'activation linéaire et logistique.

Application aux données

Le réseau de neurones sélectionné est très rapide à entrainer. De plus l'AUC sur le jeu de données test est 0,9904.

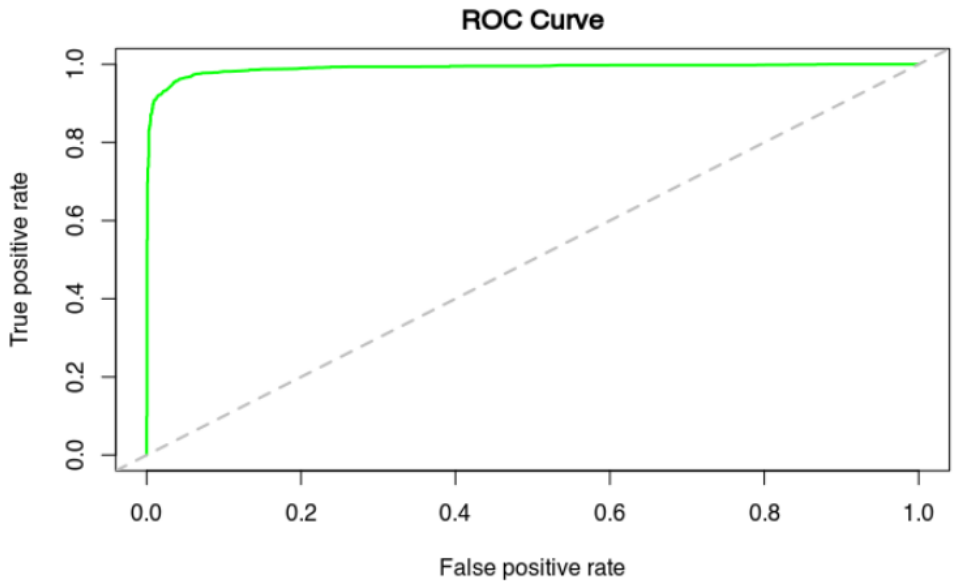


Figure 52 – Courbe ROC du Réseau de neurones sur les données test

Base d'apprentissage Quantile 0,5		Prédiction	
		0	1
Valeur	0	8124	483
	1	413	12811

Indicateurs base d'apprentissage	
Accuracy	95,90%
Taux de vrais positifs	96,88%
Taux de vrais négatifs	94,39%
Taux de faux positifs	5,61%
Précision	96,37%

Base de test Quantile 0,5		Prédiction	
		0	1
Valeur	0	2042	110
	1	111	3195

Indicateurs base test	
Accuracy	95,95%
Taux de vrais positifs	96,64%
Taux de vrais négatifs	94,89%
Taux de faux positifs	5,11%
Précision	96,67%

Figure 53 – Matrices de confusion et indicateurs de performance du réseau de neurones

L'avantage principal de ce modèle est qu'il est très rapide à optimiser sur la base de données. La précision et l'*accuracy* sont un peu plus faibles que pour *XGBOOST* mais est du même ordre que *Random Forest*. Il pourrait donc être préférable à *Random Forest*.

4.6.8 *Stacking*

Description du modèle

PRINCIPE DE L'ALGORITHME

Le *stacking* (empilement) consiste à utiliser les prédictions d'autres modèles pour constituer une prédiction plus robuste. Il a été inventé lors d'une compétition *Kaggle*.

REPRESENTATION A L'AIDE DE COUCHES

Le *stacking* utilise deux couches :

- Une première couche est constituée de différents modèles qui sont appliqués au jeu de données.
- Une deuxième couche applique un nouveau modèle en utilisant les prédictions obtenues. Les réseaux de neurones sont souvent utilisés pour cette deuxième couche.

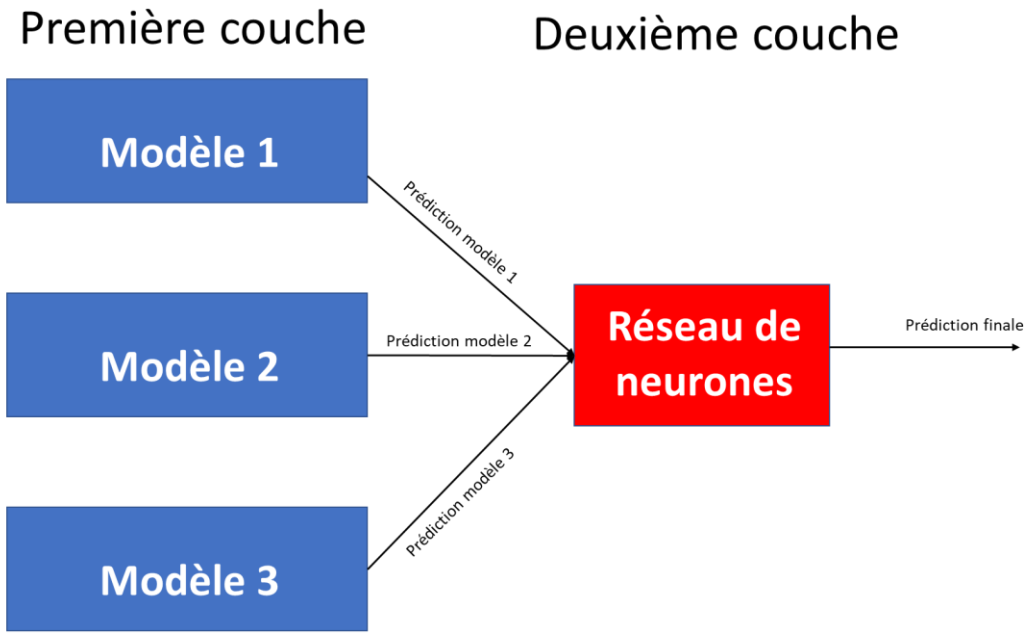


Figure 54 – Principe de fonctionnement du *stacking*

Application aux données

Les trois modèles de la première couche choisis sont :

- Un modèle *Random Forest*,
- Un modèle *XGBOOST*,
- Un réseau de neurones.

Les probabilités obtenues par les modèles sont stockées et forme 3 variables explicatives à la variable Doublon. Ensuite, un réseau de neurones (avec choix des hyperparamètres) est sélectionné pour former la deuxième couche.

Avec cette méthode, l'AUC obtenue est 0,9949.

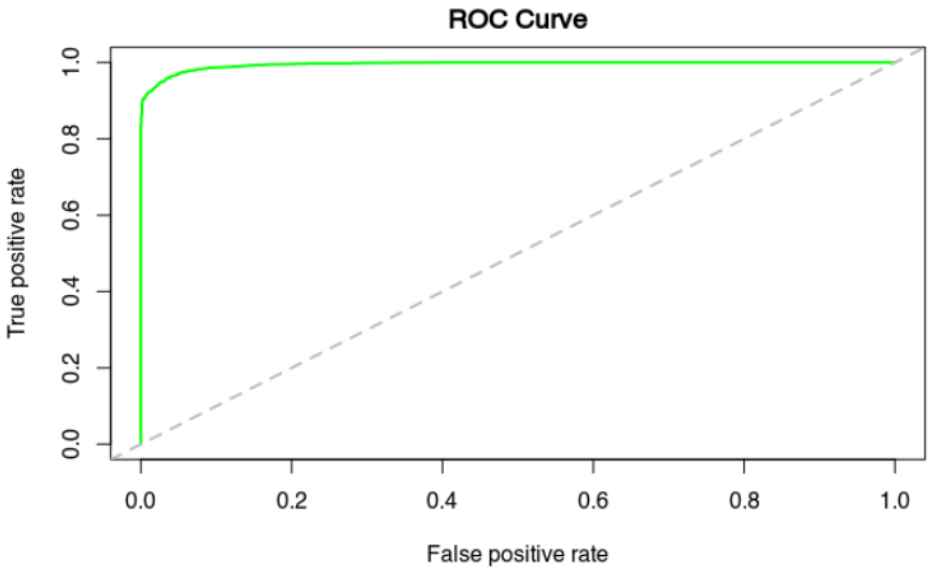


Figure 55 – Courbe ROC du *Stacking* sur les données test

Base d'apprentissage Quantile 0,5		Prédiction	
		0	1
Valeur	0	8247	360
	1	381	12843

Indicateurs base d'apprentissage	
Accuracy	96,61%
Taux de vrais positifs	97,12%
Taux de vrais négatifs	95,82%
Taux de faux positifs	4,18%
Précision	97,27%

Base de test Quantile 0,5		Prédiction	
		0	1
Valeur	0	2066	86
	1	121	3185

Indicateurs base test	
Accuracy	96,21%
Taux de vrais positifs	96,34%
Taux de vrais négatifs	96,00%
Taux de faux positifs	4,00%
Précision	97,37%

Figure 56 – Matrices de confusion et indicateurs de performance du *Stacking*

L'*accuracy* et la précision sont les meilleures de tous les modèles présentés. Cependant, l'utilisation de ce modèle requiert l'apprentissage de tous les autres modèles et est donc bien plus coûteux en temps.

4.6.9 Limites de l'étude

Le retraitement de la base de données ne peut pas être parfait. En effet, parmi les cas avec régularisation, il est impossible de savoir avec une précision parfaite quels cas concernent des doubles règlements. De plus, parmi les cas sans régularisation, il est impossible de contrôler tous ces cas. Ainsi, les règles métiers ont permis de considérer certains de ces cas comme doublons. Cependant, ce retraitement peut conduire à mal classer certaines lignes et le modèle est forcément parfait par construction.

La base de données a été construite sur un périmètre restreint. En effet, d'autres doublons de règlements existent. Certains de ces cas sont présentées dans la partie suivante.

L'algorithme construit à l'aide du réseau de neurones est désormais appliqué chaque jour sur les cas de la veille. De ce fait, le contrôle des cas de la veille par les gestionnaires permet d'effectuer des blocages de virement pour les indus. Après un test sur plusieurs semaines, le taux de positifs se situe aux alentours des 90%.

4.7 Autres types de doubles règlements non considérés

4.7.1 Bénéficiaires différents

La recherche sur les doubles règlements ont permis de relever d'autres atypies, en regardant les soins pour lesquels :

- Les dates de soins sont identiques,
- Le montant des frais réels est identique,
- Les deux soins ont été réglés et ont résulté en une prise en charge supérieure à ce qui est prévu par le contrat,
- Les soins ont été réalisés par un bénéficiaire différent,
- Les dates de soins sont postérieures au 01/07/2020.

Dans ce cadre, des règlements de lignes de soins similaires ont été observés pour l'hospitalisation.

Un premier cas a mis en lumière l'importance de bien renseigner le bénéficiaire d'un soin. Sur ce contrat, de nombreuses hospitalisations ont eu lieu pour un assuré. Cependant, en vérifiant les factures des soins rapprochées par la requête, il s'avère qu'il y a eu des erreurs récurrentes sur le bénéficiaire renseigné sur ces hospitalisations. De ce fait, deux règlements de chambre particulière ont été effectués à tort sur les deux bénéficiaires.

D'autres cas présentent des personnes sur un même contrat étant de manière récurrente en hospitalisation en même temps et pour les mêmes dates de soins en chambre particulière. Il est alors licite de se demander si ces deux personnes sont bien dans deux chambres séparées. Les cas ont été relevés et des actions vont être menées. Ces cas seront analysés par l'équipe dédiée à la lutte contre la fraude.

Ces soins identiques réglés pour deux bénéficiaires différents pourraient être dans certains cas des fraudes de la part de l'assuré qui falsifie son document en mettant le nom d'une autre personne sur le contrat. Cela peut également être une erreur à la gestion d'une saisie du bénéficiaire qui conduit à deux remboursements comme dans les cas d'hospitalisations évoqués dans cette partie.

Ces cas n'ont pas été étudiés dans un premier temps pour les raisons suivantes :

- Ces cas résultent en un nombre important de faux-semblants,
- L'analyse est plus difficile et la fraude si elle a lieu doit être prouvée,
- La nouvelle table des bénéficiaires est en cours de finalisation, il a donc fallu utiliser une table moins optimisée pour obtenir des extractions rapidement.

4.7.2 Dates de soins différentes

Pour des doubles règlements, les dates de soins peuvent être différentes. Différents cas possibles pourraient conduire à ces erreurs :

- Une erreur de saisie de la date de soins et une communication de deux documents qui peuvent conduire à la saisie de deux documents,
- Une falsification de la date du document de la part de l'assuré,
- Les soins ont été réalisés sur une période de plusieurs jours (hospitalisation) et il y a un chevauchement des périodes de soins anormal comme développé dans la partie chevauchement de chambres particulières.

Les deux premiers cas, qui conduiraient à plus de temps d'exécution des programmes de construction des bases de données, pourraient néanmoins permettre de détecter d'autres indus. Il pourrait être intéressant de construire différents indicateurs comme le temps entre deux prestations identiques pour pouvoir analyser ces différents cas.

Pour les cas d'erreurs de saisie ou de falsification de date du document, la distance de *Levenshtein* pourrait être envisageable pour rapprocher des lignes de soins proches.

4.7.3 Contrats différents

Des doubles règlements sont possibles sur deux contrats différents.

Tout d'abord la deuxième extraction évoquée sur le travail concernant les soins hors période de garantie de contrat relève une partie de ces cas.

D'autres cas sont possibles. C'est pourquoi une extraction a été réalisée pour rechercher les soins payés pour un même bénéficiaire sur deux contrats différents. La difficulté pour cette extraction est que le bénéficiaire est directement relié au décompte santé. De ce fait, rechercher les doublons au niveau de la ligne de soins est trop coûteux en termes de temps d'exécution.

Ainsi, les couples de décomptes santé avec les caractéristiques suivantes ont été remontés :

- Les deux décomptes ont les mêmes frais réels, dates de soins, bénéficiaires,
- Les décomptes ont été payés sur deux contrats différents,
- Les règlements ACM résultent tous les deux en un règlement de plus de 50 euros,
- $RembACM_1 + RembACM_2 + Max(MontantRO_1, MontantRO_2) > FraisRéels$,
- Il n'y a pas eu de régularisations pour les deux décomptes,
- La date de soins commune est supérieure au 01/01/2022.

Cette extraction a pour but de trouver les décomptes identiques réglées sur deux contrats différents sans régularisation pour lesquels l'assuré a perçu plus que les frais réels.

Cela peut conduire à trouver différents types de remboursements à tort :

- Un des deux contrats est en surcomplémentaire de l'autre contrat et le remboursement a été mal effectué,
- Un même assuré est présent sur deux contrats différents suite à une fraude (multi-assurance) ou à une erreur de gestion et il y a donc un paiement en double.

4.8 Estimation du risque annuel sur le périmètre de détection

Le but de cette partie est d'estimer le risque lié aux doubles règlements non détectés du périmètre choisi. Pour cela, des lois de probabilité vont être choisies pour modéliser le nombre potentiel de règlements en double et leur impact unitaire.

À l'aide de l'algorithme obtenu avec les réseaux de neurones, une base de données des doubles règlements potentiels est construite pour les décomptes saisis entre 2018 et 2022 sans régularisation. L'apprentissage a été effectué au niveau des lignes de soins. Pour quantifier le risque, le choix est de se placer au niveau du décompte de soins. Ainsi, en sommant les règlements des lignes de soins présumées payées en double par décompte, une base de doubles règlements potentiels non détectés est construite. Le tableau suivant est obtenu :

Année	Nombre de doubles règlements
2018	1129
2019	1541
2020	1479
2021	1358

Figure 57 – Nombre de décomptes avec des règlements en double potentiels saisis par année

L'année 2018 est exclue du calcul de risque. De même, 2022 n'est pas prise en compte car les soins liés à une période d'hospitalisation se comportent de manière particulière. S'il y a une erreur dans une facture, l'hôpital ou la clinique émet une nouvelle facture et prend contact plus tard avec l'assurance pour rembourser l'indu lié à la première facture. Ainsi, une partie des doubles règlements non régularisés pour 2022 peuvent être dus à cette procédure car ils seront régularisés plus tard.

En prenant en compte les décomptes saisis entre 2019 et 2021, les statistiques suivantes sont obtenues :

Statistiques	
Moyenne	1 459
Ecart-type	93
Variance	8 662
Variance / Moyenne	5,94

Figure 58 – Statistiques concernant le nombre de décomptes avec des règlements en double potentiels saisis par an

$\frac{Variance}{Moyenne} \gg 1$ donc la loi binomiale négative est choisie pour modéliser le nombre de décomptes avec des règlements en double saisis par an.

Soit $N \sim \mathcal{BN}(n, p)$. Alors, les moments de N sont donnés par :

$$\mathbb{E}[N] = \frac{n \times (1-p)}{p}$$

$$\mathbb{V}(N) = \frac{n \times (1-p)}{p^2}$$

En utilisant la méthode des moments, les paramètres de la loi sont obtenus.

Reste désormais à sélectionner une loi pour le montant de règlement en double unitaire par décompte noté DR. Pour cela, les packages *fitdistrplus* et *AdequacyModel* sont utilisés. Le périmètre choisi contient des doubles règlements pour des paiements supérieurs à 50€. Pour se ramener plus facilement à des lois de probabilités usuelles, l'adéquation de loi de probabilité est effectuée pour $DR - 49{,}9$.

La densité empirique et la fonction de répartition empirique sont tracés :

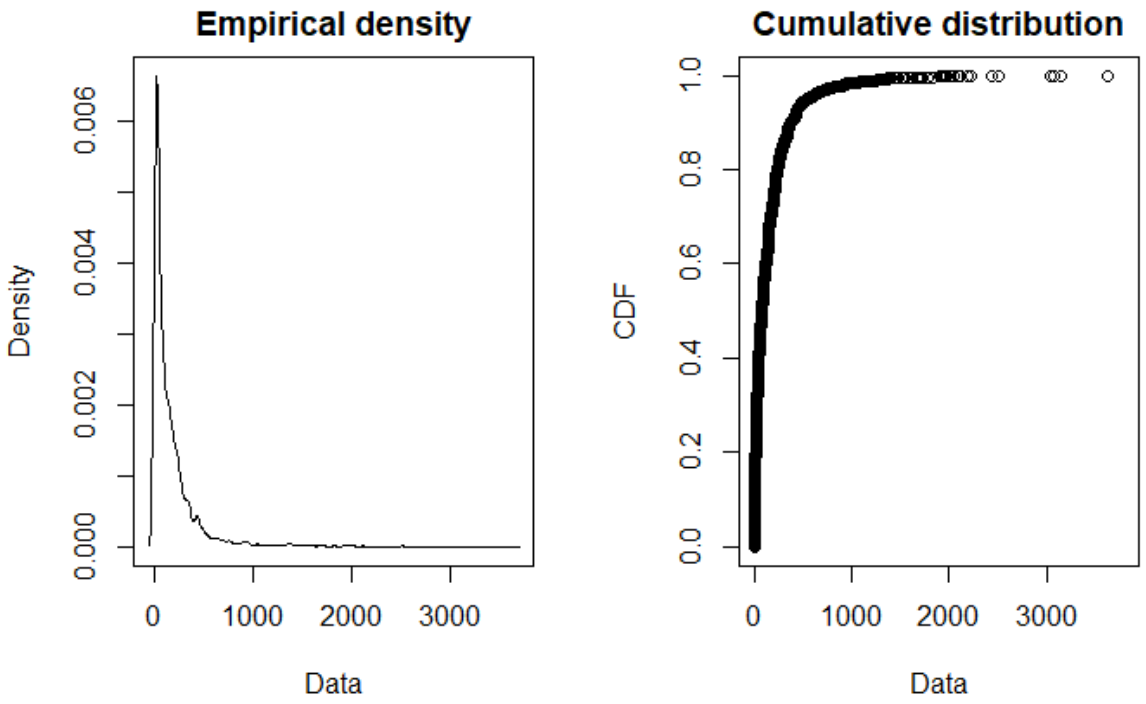


Figure 59 – Densité et fonction de répartition empiriques des indus liés aux règlements en double

D'après la forme de la densité, les lois suivantes sont envisagées :

- Loi de Weibull,
- Loi Gamma,
- Loi Log-normale,
- Loi de Pareto.

Pour mesurer graphiquement l'adéquation des lois, différents graphiques sont utilisés :

- La comparaison entre l'histogramme des données et les densités théoriques,
- La comparaison entre la fonction de répartition empirique et les fonctions de répartitions théoriques des lois envisagées,
- La comparaison entre les quantiles empiriques et les quantiles des lois envisagées,
- La comparaison entre les probabilités empiriques et les probabilités des lois envisagées.

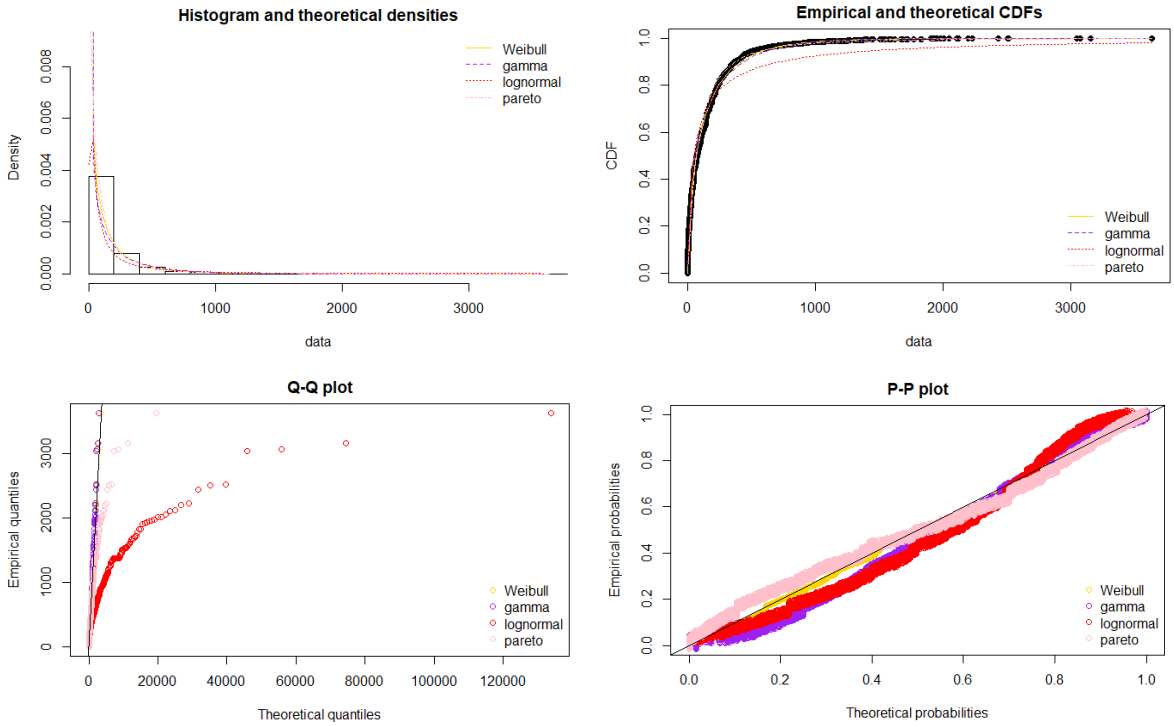


Figure 60 – Graphiques de comparaison entre les données empiriques et les lois de probabilité sélectionnées

Sur le graphique comparant les fonctions de répartition, la loi log-normale semble être la moins adaptée aux données. Il en est de même sur le QQ Plot concernant la loi log-normale et la loi de Pareto. La loi de Weibull semble être celle qui convient le mieux aux données.

Différentes statistiques peuvent être utilisées pour mesurer l'adéquation des lois de probabilités :

Statistic	General formula	Computational formula
Kolmogorov-Smirnov (KS)	$\sup \lvert F_n(x) - F(x) \rvert$	$\max(D^+, D^-)$ with $D^+ = \max_{i=1,\dots,n} \left(\frac{i}{n} - F_i\right)$ $D^- = \max_{i=1,\dots,n} \left(F_i - \frac{i-1}{n}\right)$
Cramer-von Mises (CvM)	$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum_{i=1}^{n} \left(F_i - \frac{2i-1}{2n}\right)^2$
Anderson-Darling (AD)	$n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dx$	$-n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \log(F_i(1 - F_{n+1-i}))$

Figure 61 – Statistiques pour comparer l'adéquation des lois de probabilité

Le package *fitdistrplus* et la fonction *gofstat* permettent d'obtenir les statistiques évoquées :

```
Goodness-of-fit statistics
                                 weibull       gamma    lognormal      pareto
Kolmogorov-Smirnov statistic  0.03776136   0.1031462   0.09475285  0.06626424
Cramer-von Mises statistic    1.29482377  20.3800444  25.16865323  8.23873415
Anderson-Darling statistic   13.31788533 120.9965267 170.27538350 81.08059793

Goodness-of-fit criteria
                               weibull    gamma lognormal   pareto
Akaike's Information Criterion 78616.76 79006.32  80848.92 79409.23
Bayesian Information Criterion 78630.37 79019.93  80862.52 79422.83
```

Figure 62 – Valeurs des statistiques permettant de comparer l'adéquation des lois de probabilité sélectionnées aux données

Selon les statistiques obtenues, la loi de Weibull est sélectionnée.

Pour modéliser le fait qu'un décompte sélectionné peut ne pas être forcément lié à un indu, une loi de Bernoulli est utilisée. Le paramètre de la loi sera alors 90% étant donné la phase de test de l'algorithme par les gestionnaires qui conduit vers une précision de 90%.

La loi du risque est alors modélisée de la manière suivante :

$$Risque = \sum^{N} DR_k \times B_k$$

Avec :

- N une variable aléatoire de loi binomiale négative avec les paramètres obtenus qui modélise le nombre de règlements en double potentiels par an,
- $(DR_k - 49{,}9)$ une suite de variables aléatoires indépendantes identiquement distribuées de loi de Weibull avec les paramètres obtenus qui modélisent l'impact d'un double règlement potentiel,
- (B_k) une suite de variables aléatoires indépendantes identiquement distribuées de loi de Bernoulli de paramètre 90% modélisant la précision de l'algorithme,
- $N, (DR_k), (B_k)$ indépendantes.

Avec 1 000 000 de simulations, l'histogramme du risque est tracé.

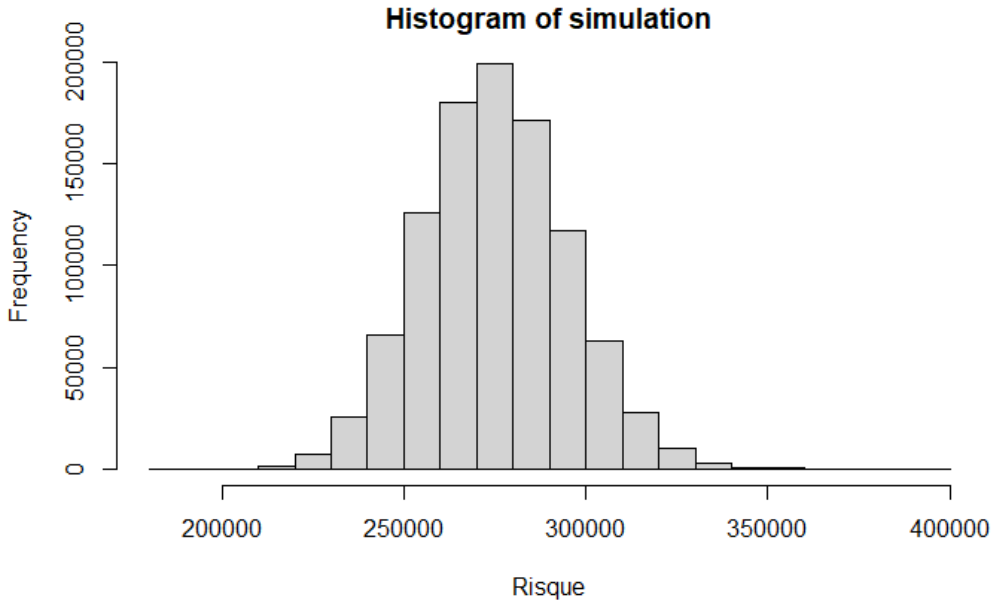


Figure 63 – Histogramme des simulations du risque

Ainsi, $E(Risque) = 275\,000$ et $VaR_{99,5\%}(Risque) = 330\,000€$.

Conclusion

Les sources de remboursements à tort sont nombreuses en assurance santé du fait du grand volume de prestations. Le but de ce rapport était d'utiliser la classification supervisée sur les doubles règlements en assurance santé.

Pour cela une base de données pour des frais réels, dates de soins, bénéficiaires et contrats identiques a d'abord été construite en prenant en compte différents critères sources de règles métiers et d'analyses effectuées à l'aide de la création d'indicateurs liés aux informations des lignes de soins potentiellement en doublon. Cette étude a permis de récupérer des indus importants sur ce périmètre de règlements en double. Elle permet chaque jour de détecter des cas de règlements en double saisis la veille pour bloquer des paiements indus.

Une fois cette base construite, les algorithmes de classification supervisée ont été appliqués. Les résultats de l'application des algorithmes semblent faire apparaître une problématique de sur-apprentissage pouvant s'expliquer par la construction de la base de données. Ce problème provient du fait qu'il est impossible de savoir avec précision si les lignes de la base de données font référence à un règlement en double ou non. Cela a conduit à utiliser des règles métiers à l'aide d'indicateurs pour construire la base. Cependant, cette construction apporte un biais au modèle car les algorithmes vont utiliser ces mêmes indicateurs. Par construction, le modèle est donc forcément un modèle parfait. La mise à disposition chaque jour aux gestionnaires des cas de règlements en double potentiels saisis la veille va permettre une meilleure application des algorithmes d'apprentissage supervisée. Il sera ainsi possible de mettre à jour les modèles avec les indus réellement topés, en supprimant petit à petit les règles métiers une fois la volumétrie suffisante.

D'autres types de doubles règlements peuvent être exploités et ont déjà donné des résultats sur de simples requêtes. Ces doubles règlements peuvent être payés sur des contrats, des bénéficiaires ou des dates de soins différentes. L'exploitation de ces cas est plus complexe mais pourrait permettre de détection des doubles règlements ayant d'autres causes (erreur à la souscription ou multi-assurance avec deux contrats différents, falsification de documents avec changement de bénéficiaire ou de date de soins).

Sur le périmètre choisi, l'adéquation de lois de probabilités au nombre d'indus potentiels et au coût unitaire permet d'estimer le montant du risque et d'en donner une *Value At Risk* à 200 ans. Cette estimation a été effectuée en prenant en compte les coûts des indus potentiels du périmètre choisi de 2018 à 2022 et le nombre d'indus potentiels de 2019 à 2021. L'estimation du risque est limitée par les évolutions du portefeuille et des méthodes de détections au fil des années qui peuvent faire varier les indus. C'est pour cela qu'un nombre d'années limité a été choisi.

Bibliographie

Articles

BREIMAN, L. (1996, août 1). *Bagging predictors*. Machine Learning. 24, 123–140. **https://doi.org/10.1007/BF00058655**

BREIMAN, L. (2001). *Random Forests*. Machine Learning, 45, 5–32. **http://dx.doi.org/10.1023/A:1010933404324**

CHEN, T., & GUESTRIN, C. (2016). *XGBoost : A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA : ACM. **https://doi.org/10.1145/2939672.2939785**

Classification Commune des Actes Médicaux Guide de lecture et de codage. (2008). Agence Technique de l'Information sur l'Hospitalisation. **https://www.atih.sante.fr/sites/default/files/public/content/1678/guide_lecture_complet_01082008.pdf**

DELIGNETTE-MULLER, M. & DUTANG, C. (2015). *fitdistrplus : An R Package for Fitting Distributions*. Journal of Statistical Software. Volume 64, Issue 4. **https://www.jstatsoft.org/index.php/jss/article/view/v064i04/2989**

FRIEDMAN, J. (2002). *Computational Statistics & Data Analysis*. 38, 367–378. **https://doi.org/10.1016/S0167-9473(01)00065-2**

GUARINO, L. (2022). *« Le potentiel de fraude en Europe est estimé à 10 % des prestations payées »*. La Tribune de l'Assurance. **https://tribune-**

assurance.optionfinance.fr/lessentiel/le-potentiel-de-fraude-en-europe-est-estime-a-10-des-prestations-payees.html

Rédaction, L. A. (2017). *Qu'advient-il des indemnités versées à l'assuré à la suite d'une fraude ?* Argus de l'Assurance. **https://www.argusdelassurance.com/acteurs/qu-advient-il-des-indemnites-versees-a-l-assure-a-la-suite-d-une-fraude.118535**

Codes juridiques

Article L.113-8 du Code des Assurances.

Article L.113-9 du Code des Assurances.

Article L.114-1 du Code des Assurances.

Article L.322-26-1 du Code des Assurances.

Articles L.911-1 à L.911-8 du Code des Assurances.

Articles D.911-0 à D.911-8 du Code des Assurances.

Article 1235 du Code Civil.

Article 1376 du Code Civil.

Article 2219 du Code Civil.

Article 2714 du Code Civil.

Article L.111-1 du Code de la Mutualité.

Article L.114-16 du Code de la Mutualité.

Cours

ALBERT, M., BESSE, P., LAURENT, B., & ROUSTANT, O. (2021). *Machine Learning*. INSA Toulouse, Cours Master Mathématiques appliquées. **https://www.math.univ-toulouse.fr/~besse/pub/machineLearning.pdf**

BIRMELE, E. (2020). *Apprentissage supervisé*. Université de Strasbourg, Cours Master Actuariat. **https://helios2.mi.parisdescartes.fr/~ebirmele/depots/Enseignements/Strasbourg/apprentissage.pdf**

BOISUMEAU, N. (2022). *Droit de l'assurance*. Université de Strasbourg, Cours Master 2 Actuariat.

BEIL, C. (2021). *Assurance dépendance*. Université de Strasbourg, Cours Master 1 Actuariat.

BERARD, J. (2021). *Modèles linéaires généralisés*. Université de Strasbourg, Cours Master 1 Actuariat.

LALLEMENT, T. (2022). *Provisionnement non-vie*. Université de Strasbourg, Cours Master 2 Actuariat.

MAURICE, B. (2018). *Fonctionnement du neurone artificiel*. Deeply Learning. **https://deeplylearning.fr/cours-theoriques-deep-learning/fonctionnement-du-neurone-artificiel/**

ROUVIERE, L. (2019). *Machine learning*. Université de Rennes 2, Cours. **https://lrouviere.github.io/ml_lecture/**

SAUGET, M. (2008). *Parallélisation de problèmes d'apprentissage par des réseaux neuronaux artificiels. Application en radiothérapie externe*. Université de Franche-Comté. **https://tel.archives-ouvertes.fr/tel-00260013/document**

TABOPDA WAFO, V. (2022). *Comparaison des méthodes de sélection des variables appliquées à la tarification des produits d'assurance non-vie*. Faculté des sciences, Université catholique de Louvain, Master en sciences actuarielles. **https://dial.uclouvain.be/downloader/downloader.php?pid=thesis%3A33671&datastream=PDF_01&cover=cover-mem**

THOME, N., FERECATU, M., AUDEBERT, N., & CRUCIANU, M. (2016). *Apprentissage statistique : modélisation décisionnelle et apprentissage profond*. CNAM, UE RCP209. **http://cedric.cnam.fr/vertigo/cours/ml2/index.html**

Livres

BOCQUAIRE, E., CHARLES, N., & MILLOT, R. (2017). *Pratique de l'assurance santé* (4e éd.). L'ARGUS de l'assurance.

COUILBAULT, F. (2015). *Les grands principes de l'assurance* (12e éd.). L'ARGUS de l'asssurance - Les Fondamentaux.

DE BOISSIEU, J. (2005). *Introduction à l'assurance - acteurs, marché, contrats, technique*. L'ARGUS de l'assurance.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning : with Applications in R*. Springer.

Mémoires

BOSQUILLON DE JENLIS, C. (2016). *La fraude au niveau des remboursements des frais de santé d'un organisme complémentaire*. Mémoire d'actuariat, CEA. **https://www.institutdesactuaires.com/docs/mem/8d9b423b84f728e1f794768c7af83d32.pdf**

HULOT, D. (2019). *Implémentation d'un modèle de détection de fraude à l'assurance dans le cadre de soins hospitaliers*. Mémoire d'actuaire, ENSAE. **https://www.institutdesactuaires.com/docs/mem/a57dd001e5a0259605f3bc4d9436cafa.pdf**

MANICKAM, S. (2020). *Modélisation de la lutte anti-fraude en santé optique à l'aide de techniques d'intelligence artificielle*. Mémoire d'actuariat, ISFA. **https://www.institutdesactuaires.com/docs/mem/afa5097ee0e0424468541dab6f1e70c2.pdf**

PALIS, M. (2021). *Impact de la réforme du « 100 % Santé » sur les contrats individuels et collectifs de complémentaires santé*. Mémoire d'actuaire, ISFA. **https://www.institutdesactuaires.com/docs/mem/a6c31300e3cae07163cda71a90b5db45.pdf**

Packages du logiciel R

ARNOLD, J. (2021). *Package 'ggthemes' : Extra Themes, Scales and Geoms for « ggplot2 »*. Version 4.2.4. **https://cran.r-project.org/web/packages/ggthemes/ggthemes.pdf**

DELIGNETTE-MULLER, M. & DUTANG, C. (2022). *Package 'fitdistrplus'*. Help to Fit of a Parametric Distribution to Non-Censored or Censored Data. Version 1.1-8. **https://cran.r-project.org/web/packages/fitdistrplus/fitdistrplus.pdf**

DINIZ MARINHO, P., BOURGUIGNON, M. & BARROS DIAS, C. (2022). *Package 'AdequacyModel'*. Adequacy of Probabilistic Models and General Purpose Optimization. Version 2.0.0. **https://cran.r-project.org/web/packages/AdequacyModel/AdequacyModel.pdf**

GOULET, V., DUTANG, C., PIGEON, M., A. RYAN, J., GENTLEMAN, R., IHAKA, R., R CORE TEAM & R FOUNDATION. (2022). *Package 'actuar'*. Actuarial Functions and Heavy Tailed Distributions. Version 3.3-1. **https://cran.r-project.org/web/packages/actuar/actuar.pdf**

KUHN, M. (2022). *Package 'caret' : Classification and Regression Training*. Version 6.0-93. **https://cran.r-project.org/web/packages/caret/caret.pdf**

MEYER, D., DIMITRIADOU, E., HORNIK, K., WEINGESSEL, A., & LEISCH, F. (2022). *Package 'e1071' : Misc Functions of the Department of Statistics, Probability Theory Group (Formerly : E1071), TU Wien*. Version 1.7-11. **https://cran.r-project.org/web/packages/e1071/e1071.pdf**

ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J., & MULLER, M. (2021). *Package 'pROC' : Display and Analyze ROC Curves*. Version 1.18.0. **https://cran.r-project.org/web/packages/pROC/pROC.pdf**

SING, T., SANDER, O., BEERENWINKEL, N., & LENGAUER, T. (2020). *Package 'ROCR' : Visualizing the Performance of Scoring Classifiers*. Version 1.0-11. **https://cran.r-project.org/web/packages/ROCR/ROCR.pdf**

WEI, T., & SIMKO, V. (2021). *Package 'corrplot' : Visualization of a Correlation Matrix*. Version 0.92. **https://cran.r-project.org/web/packages/corrplot/index.html**

WICKHAM, H., CHANG, W., HENRY, L., LIN PEDERSEN, T., TAKAHASHI, K., WILKE, C., WOO, K., YUTANI, H., & DUNNINGTON, D. (2022). *Package 'ggplot2' : Create Elegant Data Visualisations Using the Grammar of Graphics*. Version 3.3.6. **https://cran.r-project.org/web/packages/ggplot2/index.html**

WICKHAM, H., FRANCOIS, R., HENRY, L., & MULLER, K. (2022). *Package 'dplyr' : A Grammar of Data Manipulation*. Version 1.0.9. **https://cran.r-project.org/web/packages/dplyr/dplyr.pdf**

Webographie

Assurance Maladie. (2022). Notre environnement : la Sécurité sociale. **https://assurance-maladie.ameli.fr/qui-sommes-nous/organisation/securite-sociale**

Les honoraires de dispensation en officine. (2020). Caisse de Prévoyance et de Retraite du personnel de la SNCF. **https://www.cprpsncf.fr/les-honoraires-de-dispensation-en-officine**

Table des figures