République Tunisienne
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Carthage - Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Rapport de Projet de Fin d'Etudes soumis afin d'obtenir le

Diplôme National d'Ingénieur en Statistique et Analyse de l'Information

Réalisé par

Ali Belkhir

Estimation des provisions pour sinistres à payer par GLM et détection des comportements atypiques

---

Soutenu le 12/06/2023 devant le Jury composé de :

Pr. Mokhtar KOUKI	Président
Dr. Farouk MHAMDI	Rapporteur
Mr. Mahmoud ALJAN	Encadrant entreprise
Dr. Ines ABDELJAOUED TEJ	Encadrante académique

Projet de Fin d'Etudes fait à

Bridges

Remerciements

En guise de préambule je tiens à adresser mes remerciements à toutes les personnes grâce auxquelles ce travail a été rendu possible.

Je souhaite exprimer ma gratitude à Mr. Mahmoud ALJAN, pour m'avoir donné envie de réaliser un mémoire sur le calcul des provisions pour sinistres et la détection des comportements atypiques au sein de « Bridges S.A ». Je le remercie également pour son accueil chaleureux à chaque fois que j'ai sollicité son aide, ainsi que pour ses multiples encouragements. J'ai été extrêmement sensible à ses qualités humaines d'écoute et de compréhension tout au long de ce travail de mémoire.

Je tiens à remercier mon encadrante universitaire, Dr. Ines ABDELJAOUED TEJ, pour la confiance qu'elle m'a accordée en acceptant d'encadrer ce travail, pour ses multiples conseils et pour toutes les heures qu'elle a consacrées à diriger ce travail. Je suis honoré d'avoir pu bénéficier de sa supervision et je lui suis extrêmement reconnaissant pour l'opportunité qui m'a été offerte.

Je remercie très chaleureusement les membres de jury qui ont accepté d'évaluer ce travail : Pr. Mokhtar KOUKI et Dr. Farouk MHAMDI qui ont consacré leur temps et leur expertise dans l'évaluation de mon projet. Leur présence et leur intérêt ont renforcé ma motivation à réaliser ce travail avec rigueur et excellence.

Ma reconnaissance va à ceux qui ont plus particulièrement assuré le soutien affectif de ce travail : ma famille ainsi que mes amis.

Résumé

La fraude et le provisionnement représentent deux enjeux majeurs en assurance non-vie, en particulier en assurance maladie. Nous nous sommes intéressés, dans ce projet, à résoudre ces deux problèmes.

Les provisions sont, généralement, calculés à l'aide des méthodes Chain Ladder. Ces méthodes nécessitent la vérification de certaines hypothèses fortes. L'un des objectifs de ce projet est de mettre en place une approche robuste pour le remplissage du triangle de liquidation, ainsi que le calcul des provisions pour sinistres à payer (PSAP) à l'aide d'un modèle GLM.

Notre deuxième objectif est de procéder à un processus de détection de fraude via l'apprentissage non-supervisé. Pour se faire, nous nous sommes intéressé au modèle de détection d'anomalie Isolation Forest. Ce modèle a pour fonction d'isoler les observations aberrantes de l'ensemble de données dans un espace multidimensionnel. Ce modèle est défini comme étant une boîte noir (Black Box). Pour cette raison, nous avons recours à l'Explainable AI (XAI) par l'approche SHAP qui nous renseigne sur la contribution des différentes variables à chaque prédiction. Une étape finale consiste à regrouper les anomalies selon leurs tendance afin d'identifier les fraudeurs et les non-fraudeurs.

L'utilisation des bibliothèques avancées du langage Python nous a facilité la réalisation des différentes tâches de ce projet.

Mots clés: ***Assurance maladie, Provisionnement, Chain Ladder, GLM, Fraude, Apprentissage non-supervisé, Isolation Forest, Explainable AI, SHAP, K-means***

Abstract

Fraud and reserving represent two major issues in non-life insurance, particularly health insurance. In this project, we set out to solve these two problems.

Provisions are generally calculated using Chain Ladder methods. These methods require the verification of certain strong assumptions. One of the aims of this project is to develop a robust approach to filling the liquidation triangle, and to calculating the reserves needed for IBNR (Incurred But Not Reported), using a GLM model.

Our second objective is to develop a fraud detection process using unsupervised learning. To deal with, we focused on the anomaly detection model : Isolation Forest. The model's role is to isolate outliers from the data set in a multidimensional space. This model is defined as a Black Box. For this reason, we used the Explainable AI (XAI) SHAP approach, which provides information on the contribution of different variables to each prediction. The final step is to group anomalies according to their tendency, in order to identify fraudsters and non-fraudsters.

The use of advanced libraries in the Python language helped us perform these tasks.

Keywords: Health Insurance, Reserving, Chain Ladder, GLM, Fraud detection, Unsupervised Learning, Isolation Forest, Explainable AI, SHAP, K-means

Table des matières

Table des figures

Liste des tableaux

Introduction

L'assurance est un contrat entre un individu ou assuré et une compagnie d'assurance ou assureur, dans lequel l'assureur accepte de dédommager l'assuré pour toute perte spécifiée en contrepartie du versement d'une prime. Le but de l'assurance est de fournir une protection financière contre des événements inattendus, tels que le décès, les blessures, les dommages matériels ou la perte de revenus. La police d'assurance décrit les conditions et les circonstances dans lesquelles l'assureur versera une indemnité [3].

Ces entreprises et compagnies d'assurance fonctionnent de manière particulière en raison de leur gestion spécifique adaptée à leur activité principale, qui consiste à prévoir les risques afin de les assurer.

L'assurance est dominée par la loi des grands nombres [21]. Cette loi prend en compte les facteurs constants et récurrents qui conduisent typiquement à un événement imprévisible et accidentel. Ces évènements sont appelés sinistres. Malgré les avantages potentiels en théorie, la loi des grands nombres peut ne pas se manifester systématiquement dans les cas pratiques [21]. Dans certaines situations, le comportement de l'assuré peut avoir un impact sur le niveau de risque [1]. Non seulement il peut se désintéresser de la réalisation du risque, mais aussi s'intéresser à cette réalisation parce qu'elle peut lui profiter [20].

Des études sont réalisées régulièrement, afin d'estimer les dommages causés par la fraude envers les compagnies d'assurance. D'après la fédération des sociétés d'assurance européenne «Insurance Europe», le coût de la fraude en Europe est estimé à 10% du montant de la sinistralité globale. Cette proportion est plus élevée dans certaines régions du monde où le contrôle des assurances est moins

1. Voir www.investirsorcier.com/la-loi-des-grands-nombres-dans-lindustrie-des-assurances-vue-densemble/

rigoureux, notamment en Afrique, dans certaines parties de l'Asie et en Amérique du Sud [4].

D'après la Fédération Tunisienne des Sociétés d'Assurance (FTUSA), les estimations montrent que les coûts de la fraude se situent entre 5% et 10% des montants d'indemnisation. Le secteur des assurances enregistre des pertes, engendrés par la fraude à l'assurance, de 150 MDT par an [2].

Les objectifs de ce projet sont, dans un premier temps, de fournir des méthodes de calcul des réserves pour les assurances maladie, de comprendre pourquoi certaines méthodes sont inadaptées et de proposer une approche robuste pour modéliser au mieux les provisions. Dans un deuxième temps, il s'agira de détecter les consommations aberrantes en matière de sinistre et les comportements atypiques des assurés via une approche d'apprentissage statistique non supervisée.

Le mémoire s'articule selon le plan suivant : dans un premier temps, nous allons présenter le cadre général de l'étude. Plus précisément, nous tenterons de comprendre le paysage de l'assurance maladie en Tunisie et d'examiner les différentes interactions entre ses acteurs en parlant des enjeux de ce mémoire. Dans une deuxième partie, nous proposerons une alternative, par les modèles GLM, pour le calcul des provisions en cas d'invalidité des hypothèse du modèle classique (Chain Ladder), dont nous allons détailler des fondements mathématiques et présenter les résultats. Une dernière partie va être dédiée à la détection des fraudes. Plus spécifiquement, un algorithme d'apprentissage non-supervisé dit *Isolation Forest* va être exploiter pour détecter les aberrations. Cette boite noir va être décortiquée via une approche d'explication des algorithmes d'intelligence artificielle dite SHAP. finalement, en terminera par regrouper ces aberrations par l'algorithme d'apprentissage non-supervisé *K-means*.

2. Voir www.businessnews.com.tn/tunisie-les-fraudes-a-lassurance-coutent-annuellement-150-md,520,53114,3

Première partie

Cadre du projet

Chapitre 1

Problématique

Avant d'entrer dans le vif du problème, il est important d'avoir une vision claire sur l'environnement autour de l'assurance maladie en Tunisie. Ce chapitre sera donc une présentation en plusieurs points des différents acteurs du secteur de l'assurance.

1.1 Structure du système de santé en Tunisie

Bien que le système de santé tunisien ait atteint un niveau de développement élevé en Afrique et corresponde aux normes médicales de plusieurs pays européens, des disparités significatives subsistent en termes d'équité et de qualité des soins entre les régions rurales et les régions urbaines, ainsi qu'entre les secteurs public et privé.

Afin de s'attaquer à des problèmes tels que les inégalités importantes, la couverture médicale insuffisante et l'augmentation considérable des dépenses liées aux soins de santé et aux frais médicaux supportés par les ménages, la Tunisie a mis en œuvre une réforme de l'assurance maladie en 2004.

1.1.1 Système de santé public en Tunisie

Le système de santé public en Tunisie est géré par la Caisse Nationale d'assurance Maladie (CNAM). Les missions de la CNAM portent sur :

- L'assurance maladie en Tunisie est géré par la CNAM
- Administration des programmes d'indemnisation statutaires pour les accidents du travail et les maladies professionnelles dans les domaines public et privé.

- ✯ L'administration des divers régimes légaux d'assurance maladie.
- ✯ L'octroi des prestations de maladie et de maternité fournies par les régimes de sécurité sociale.

La protection sociale, gérée par la CNAM couvre les prestations de santé des tunisiens et des résidents en Tunisie, qu'ils soient retraités, travailleurs indépendants ou encore salariés.

Il est obligatoire pour les employeurs d'affilier leurs employés à la Caisse Nationale de la Sécurité Sociale (CNSS) dans un délai d'un mois à compter de leur établissement. L'immatriculation auprès de la CNAM est obligatoire pour bénéficier de la couverture médicale, d'où l'importance de vérification de de cette tache.

Les enfants mineurs et le conjoint de l'assuré bénéficient de la sécurité sociale, donc de l'assurance maladie. Tous les citoyens et résidents tunisiens ont droit à des soins de santé gratuits dans les établissements de santé et les hôpitaux publics. Si un assuré se rend dans un centre de soins de santé primaires, un hôpital de district ou un hôpital régional, les frais engagés dans les établissements publics sont remboursés.
La Caisse Nationale d'Assurance Maladie (CNAM) est chargée d'assurer la couverture de l'assurance médicale en Tunisie. Créée en 2004 dans le cadre des réformes visant à consolider le système d'assurance maladie et les prestations médicales précédemment fournies par la Caisse Nationale de Sécurité Sociale (CNSS) et la Caisse Nationale de Retraite et de Prévoyance sociale (CNRPS). L'un des principaux objectifs de la réforme de 2004 était d'améliorer la couverture d'assurance en réduisant les dépenses directes des ménages et en leur permettant de se faire soigner sans être gênés par des obstacles financiers.

Il existe, actuellement, deux principaux systèmes de mutualisation en Tunisie : la Caisse nationale d'assurance maladie (CNAM), qui fournit une couverture aux travailleurs du secteur formel, et les systèmes d'assistance médicale gratuite (AMG), qui offrent une couverture aux populations vulnérables. La couverture CNAM a trois options : la filière publique (54% des assurés sociaux de la CNAM en 2015), la filière privée ($18,5\%$ des assurés en 2015) et le système de remboursement de frais (21% des assurés en 2015) [1].

1. Voir https ://www.emro.who.int/

Le concept de solidarité apporte une première validation à la régulation du système d'assurance maladie primaire par le gouvernement. Sur le plan juridique, la CNAM fonctionne comme un établissement public, autrement dit une personne morale de droit public chargée d'assurer un service public. Cependant, le principe assuranciel fait que cette même CNAM n'est pas une institution totalement désintéressée financièrement et destinée uniquement à l'idée de servir le public. Depuis 2004, les autorités considèrent la CNAM comme une entreprise publique. Cela fait que la CNAM comme un organisme public autonome, jouissant de la personnalité juridique et de l'indépendance financière. [2]

1.1.2 Système de santé privé en Tunisie et organismes complémentaires

En Tunisie, le domaine de la santé privée bénéficie d'un développement notable, tant en termes d'infrastructures que de capacité d'accueil et de professionnels de la santé. Certaines spécialités, telles que la dentisterie et l'optique, sont principalement disponibles dans le secteur privé. 6000 lits, dans une centaine de cliniques (109 en 2021) réparties sur l'ensemble des régions, représente 20% de la capacité nationale. [3]

La complémentaire santé a pour objet de compléter les prestations de la sécurité sociale en matière de frais médicaux. Elle couvre toutes les prestations incluses dans la couverture de la sécurité sociale, telles que la consultation médicale, l'analyse médicale, l'hospitalisation, la maternité, l'ophtalmologie, les soins dentaires, l'appareillage, les médicaments et la prophylaxie.
Le secteur de santé tunisien présente des limites :

- ✯ Environ 17% des ménages ne sont pas couverts et sont contraints de supporter directement les coûts des services de santé fournis par des prestataires publics et privés. [6].
- ✯ Les cliniques et les hôpitaux privés sont plus recommandés, spécifiquement pour les traitements médicaux spécialisés. Bien que tels traitements coûtent plus chers dans les organismes de santé privé. Ainsi que ces cliniques se sitent principalement dans les grandes zones urbaines. [4].

2. Voir https ://www.cnam.nat.tn/
3. Source : l'Institut National de la Statistique (INS)
4. Voir https ://www.expatassure.com/fr/tunisie/

1.2 Problématique et cadre du projet

L'assurance collective ou contrat d'assurance de groupe est un avantage social pour les salariés fourni par l'employeur. L'employeur souscrit un contrat d'assurance (sur lequel il cotise tout ou en partie) au bénéfice de ses employés ainsi que leurs ayant droit.
Le concept de l'assurance collective réside dans le fait que :

- ➟ Le compte est ouvert pour un groupe de personnes, et non pas individuellement.
- ➟ L'employeur cotise tout ou partie au bénéfice du groupe.

En pratique, un contrat d'assurance de groupe implique la participation de trois entités distinctes :

- ➢ Le salarié assuré seul ou le salarié et ses ayants droit, soient les bénéficiaires.
- ➢ L'employeur est celui qui prend en charge de payer les primes et de signer le contrat.
- ➢ L'assurance gère les prestations et perçoit en contrepartie les primes.

La fraude à l'assurance est une préoccupation majeure pour les assureurs et peut entraîner des pertes financières considérables. Une étude récente menée par l'*Association of Certified Fraud Examiners* (ACFE) a révélé qu'au niveau mondial, les entreprises perdent environ 5% de leur chiffre d'affaires annuel à cause de la fraude. En outre, les compagnies d'assurance sont souvent la cible des fraudeurs en raison de la complexité de leurs procédures de gestion des sinistres.
La fraude oblige les assureurs à payer des indemnités pour des sinistres à de faux montants bien supérieurs aux montants réels des sinistres ou même à indemniser des sinistres n'ayant jamais eu lieu. Cela à pour effet de diminuer les capacités d'investissement des assureurs et d'obliger les compagnies d'assurance à provisionner plus que leur sinistralité réelle. Détecter les fraudes permettrait aux assureurs de doter de plus faibles provisions.

1.2.1 Organisme d'accueil : *Bridges S.A*

Bridges S.A était créée en 2020 par M. Mahmoud Aljan, son Directeur Général, et M. Ramzi Ameiri, son Directeur Technique. Elle avait obtenu le label *Startup* en mars 2021.
Bridges S.A, leader dans la gestion des sinistres santé, un acteur majeur de l'assurance et des services financiers. C'est une solution de digitalisation pour la gestion de la branche santé de l'assurance. Elle

offre à ses clients une plateforme de gestion de contrats d'assurance maladie. Cette plateforme est une solution intelligente, permet à l'assureur la gestion de son portefeuille d'assurance maladie, et aux clients le suivi de leurs soldes. Elle offre aussi une carte d'assurance maladie, cette carte offre le service tiers payant c'est-à-dire le bénéficiaire n'avance pas la partie prise en charge par l'assureur.
Aujourd'hui, après deux ans d'activité, *Bridges S.A* gère le plus grand portefeuille dans le secteur de l'assurance maladie. La stratégie de la startup s'oriente vers une satisfaction, une confiance client, une innovation et industrialisation des processus.

Bridges S.A est un partenaire fiable des compagnies d'assurance, entreprises et professionnels de santé, qui fournit des services de haute qualité respectant les normes les plus strictes en matière de transparence et de professionnalisme.
Elle soutient également les assureurs et les professionnels de santé grâce à un ensemble de technologies facilement accessibles en vue de leurs permettre de mieux communiquer avec les membres assurés pour qu'ils puissent, facilement, accéder aux soins de santé. *Bridges S.A* offre à ses membres assurés l'accès à un réseau large de professionnels de santé pour profiter des services de soins médicaux.

1.2.2 Cycle de vie d'un sinistre

Un sinistre est un évènement incertain qui entraîne des dommages et des pertes susceptibles d'être indemnisés. des tels évènements, ainsi que leur probabilité d'occurrence, représentent ce que l'on appelle risques. La compagnie d'assurance accepte d'assumer le risque qu'un assuré cherche à réduire, en échange du paiement d'une prime.

Nous allons présenter dans cette section le cycle de vie d'un sinistre i.e. la représentation de la liquidation des sinistres. Le processus d'indemnisation s'étale dans le temps.

La Figure 1.1 décrit le schéma de vie d'un sinistre quelconque :

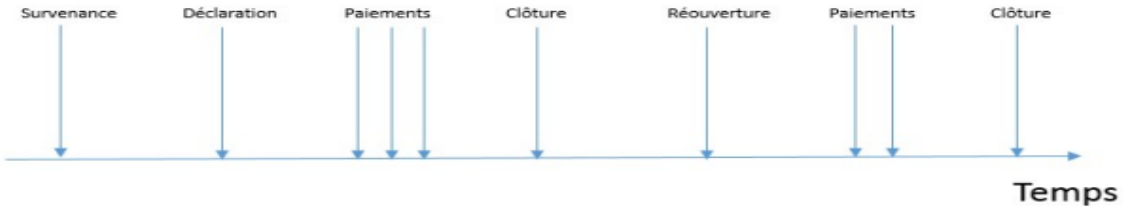


FIGURE 1.1 – Cycle de vie d'un sinistre [15]

Du point de vue de l'assureur, il existe différentes phases dans le cycle de vie d'un sinistre :

- ☞ La survenance du sinistre indique le début de son cycle de vie.
- ☞ Il peut y avoir une phase de latence en la date de survenance et la déclaration du sinistre
- ☞ Après avoir déclarer, le sinistre sera évalué puir payé (cette étape peut prendre du temps).
- ☞ Le paiement du sinistre indique sa clôture.
- ☞ Après sa clôture, le sinistre peut-être rouvert suite à une réclamation ou un éveloppement inattendu.
- ☞ Ainsi, cette réouverture implique une réévaluation du sinistre et un autre paiement
- ☞ Le paiement implique la clôture du cycle.

La compagnie d'assurance doit respecter ses engagements qu'il a envers les assurés à tout moment du cycle de vie d'un sinistre.

Les sinistres clôturés : Considérant qu'aucune indemnité ne sera accordée à l'assuré, le sinistre est officiellement clôturé à la date de l'inventaire, éliminant ainsi toute responsabilité restante.

Les sinistres ouvert : Il s'agit de sinistres déclarés avant la date d'inventaire, mais dont les montants exacts à verser à l'assuré sont encore incertains.
Ces sinistres sont connus par : IBNeR pour *Incurred But Not enough Reported*.

Les sinistres tardifs : L'estimation de ces sinistres pose le plus grand défi à l'assureur car, bien qu'ils soient survenus au cours de la période de couverture, l'assureur n'a aucune idée ni sur la quantité, ni surla gravité de ces sinistres. Ces sinistre sont connus par l'acronyme IBNyR pour *Incurred But Not yet Reported*.

1.2.3 Décomposition de la charge ultime

Les charges ultimes représentent la projection ultime des sinistres. Celle que l'actuaire doit pouvoir estimer à en utilisant des méthodes statistiques.
Cette charge se décompose comme dans la Figure 1.2 :

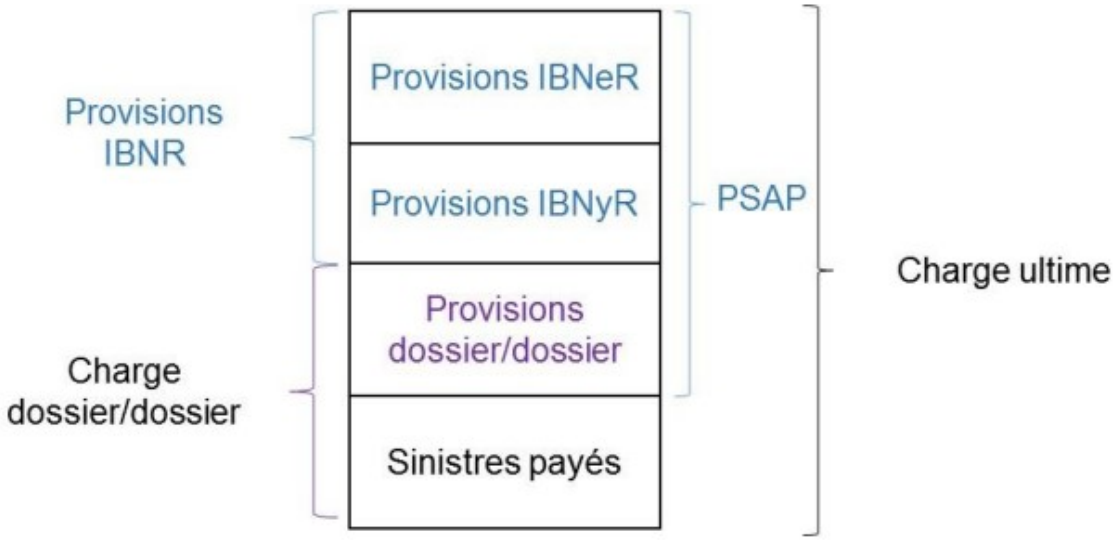


FIGURE 1.2 – Décomposition de la charge ultime [15]

On appelle **PSAP**, l'acronyme de Provision pour Sinistre À Payer par l'assureur. Ces réserves peuvent être repartis de la manière suivante :

- Les provisions dossier/dossier (D/D ou F/F en anglais pour *File/File* ou encore RBNS pour *Reported But Not Settled*). L'estimation de ces réserves est faite, cas par cas, par des gestionnaires de sinistres possédant des connaissances dans le secteur d'activité concerné.

- Les IBNR (*Incurred But Not Reported*) sont des estimations des réserves pour les sinistres potentiels qui se sont produits mais qui n'ont pas encore été déclarés à la compagnie d'assurance (IBNyR), ainsi il comprend les sinistres qui ont été déclarés mais pour lesquels les réserves constituées sont insuffisantes(IBNeR).

La charge ultime est la dépense finale correspondant au coût ultime ou définitif des sinistres. Pour

les sinistres clos, il s'agit de l'intégralité des paiements effectués par l'assureur. Pour les sinistres encore ouverts, il est calculée en tenant compte des paiements effectués jusqu'à présent, des réserves D/D déterminées par l'équipe chargée des sinistres et des IBNR estimés par les actuaires.

1.2.4 Enjeux du projet

Considérons un exemple quelconque de triangle de liquidation :

A/D	1	2	3	4	5	6
1	396	540	545	547	548	551
2	416	576	581	584	585	.
3	478	660	667	670	.	.
4	523	730	743	.	.	.
5	609	839	.	.	.	.
6	644	.	.	.	.	.

TABLE 1.1 – Exemple d'un triangle de paiements cumulés de sinistres (en MDT)

La première colonne (A) représente la date de survenance des sinistres, et la première ligne (D) correspond a la date de développement des paiements.

Nous pouvons utiliser la méthode de *Chain Ladder*[5] afin de compléter la partie manquante du triangle (Table 1.2).

5. Nous en parlerons dans le troisième chapitre de ce mémoire

A/D	1	2	3	4	5	6
1	396	540	545	547	548	551
2	416	576	581	584	585	588,20
3	478	660	667	670	671	674,86
4	523	730	743	746	747,6	751,73
5	609	839	849	849	854,3	859,02
6	644	859,02	900,1	904,1	905,7	910,64

TABLE 1.2 – Triangle complété des paiements cumulés de sinistres (en MDT)

Nous obtenons alors le montant de provision pour les sinistre à payer (PSAP) :

Année	Dernier	Ultime	PSAP
1	551	551	0
2	585	588,20	3,21
3	670	674,86	4,86
4	743	751,73	8,73
5	839	859,02	20,02
6	644	910,64	266,64

TABLE 1.3 – Calcul de Provision pour Sinistres À Payer (en MDT)

Nous calculons les PSAP comme suit :

$$PSAP_i = Ultime_i - Dernier_i \tag{1.1}$$

$$PSAP = \sum_i PSAP_i \tag{1.2}$$

Le montant de provisions est donc 303MDT.

Si nous supposons que nous détectons en moyenne 8% de fraudes (en montant),nous pouvons alors appliquer à notre triangle une détection de 8% de montants frauduleux.

De la même manière, nous complétons le triangle après détection de 8% de montants frauduleux.

A/D	1	2	3	4	5	6
1	364,32	496,80	501,40	503,24	504,16	506,92
2	382,72	529,92	534,52	537,28	538,20	541,15
3	439,76	607,20	613,64	616,40	617,49	620,87
4	481,16	671,60	683,56	686,61	687,82	691,59
5	560,28	771,88	781,12	784,61	785,99	790,30
6	592,48	818,27	828,06	831,76	833,23	837,79

TABLE 1.4 – Triangle complété des paiements cumulés de sinistres (déduits de 8% de fraudes)

Nous obtenons alors une diminution de PSAP de 8% qui s'élève à 279MDT.
Il est important de reconnaître que la méthode de calcul du PSAP utilisée ici est une méthode déterministe, ce qui signifie que le PSAP après détection de la fraude est simplement le PSAP initial avec le taux de fraude déduit.

Pour minimiser l'impact de la fraude sur leurs provisions, les compagnies d'assurance doivent mettre en place des stratégies de prévention efficaces et investir dans des technologie de détection de fraudes.

Sauf que la détection de la fraude n'est pas le seul défit pour les compagnies d'assurance, mais aussi l'estimation des provisions pour sinistres représente un autre défit pour les preneurs de décisions vu leur volatilité.

Les principaux objectifs de cette étude sont :

1. Le développement d'une approche, utilisant les modèles GLM, de calcul de Provision pour Sinistres À Payer (PSAP) en se basant sur la méthode du triangle de liquidation et la comparer aux méthodes classiques.
2. La détection des consommations atypiques en matière de sinistre en se basant sur une approche d'apprentissage statistique non supervisé.

Chapitre 2

Description de la base de données

Ce chapitre porte sur toutes les étapes de construction, de compréhension et d'analyse de l'ensemble des données utilisées dans les étapes de modélisation. Il comprend le nettoyage des données, la transformation des données pour les adapter à notre besoin, la description des variables à utiliser.

2.1 Base de données

Bridges S.A dispose d'un outil de bases de données riche en information, qui permet de conserver le plus grand nombre de données sur les personnes et les groupes assurés. Ces tables de données sont formées de plusieurs variables. Nous nous sommes intéressés à quelques unes qui sont utiles pour notre étude. Les variables les plus importantes pour nous sont présentées dans la Table 2.1.
Nous avons réalisé des transformations préalables pour adapter nos tables de données à nos besoins. Ces transformations comportent les actions suivantes :

- ✎ *Feature Selection* est une technique qui consiste à éliminer les données redondantes ou inutiles dans le problème à résoudre. Cette approche consiste à sélectionner le sous ensemble de données le plus utile pour la résolution du problème. La corrélation est une technique efficace pour la sélection.
- ✎ *Feature Construction* est une technique qui consiste à construire manuellement de nouvelles caractéristiques à partir des données brutes. Il s'agit, par exemple, de combiner des caractéristiques pour créer de nouvelles variables ou l'inverse (i.e. décomposer des variables).
- ✎ Suppression des colonnes inutiles ainsi que les colonnes qui manquent d'information.

Colonne	Description
`sous total`	Dépenses totales de l'adhérent
`pec`	Prises En Charge par l'assureur
`employeur id`	Identifiant de la police
`adherent id`	Identifiant de l'adhérent/chef de la famille
`patient id`	Identifiant du malade
`id bulletin soin`	C'est l'identifiant du bulletin de soins
`id prestation`	Identifiant de l'ensemble de bulletins de soin pour une même prestation
`date prestation`	C'est la date de prestation
`date maladie`	C'est la date de survenance du sinistre
`date heure creation`	C'est la date de payement du sinistre
`nom prestataire`	C'est le professionnel de santé (prestataire de soins)
`discipline`	C'est la discipline du professionnel de santé
`bénéficiaire qualif`	C'est la qualification du bénéficiaire
`birthday adherent`	Date de naissance du bénéficiaire

TABLE 2.1 – Description de variables les plus importantes de la base de données

2.1.1 Prétraitement pour la partie provisionnement

En ce qui concerne cette partie, nous nous contenterons des colonnes suivantes :

- `pec` : qui contient les paiements pris en charge par la compagnie d'assurance
- `date maladie` : ce sont les dates de survenance des sinistres.
- `date heure création` : ce sont les date de paiement de ces sinistres

La transformation de ces données consiste à mettre les paiements effectués sous forme d'un triangle de liquidation, que nous allons en parler ultérieurement. Cette opération se fait à l'aide d'une librairie `Python` dite `ChainLadder`.

Pour le modèle GLM

Pour appliquer le modèle GLM, nous avons besoin de transformer le triangle sous forme d'un tableau à trois colonnes. Les dates de survenance et de paiement serons des variables exogènes (explicatives) qualitatives, et une colonne contenant les montants payés sera notre variable endogène (cible).

Les variables explicatives nécessitent un prétraitement qui consiste à encoder leurs modalités. Cela se fait à l'aide de la commande `LabelEncoder()` en `Python`.

2.1.2 Prétraitement pour la partie détection des fraudes

Dans cette partie du projet, nous avons besoin du maximum d'information sur chaque adhérent, ainsi que les sinistres qui lui correspondent. Pour cela, nous allons regrouper l'ensemble de données, observées sur une année, correspondant à chaque adhérent dans une ligne (observation).

⇨ Une analyse transversale va être effectuée.

La nouvelle table de donnée va contenir les informations montrées dans la Table 2.2.

Colonne	**Description**
`adherent`	C'est l'identifiant de l'adhérent
`Age`	L'âge de l'adhérent
`nbr malade`	C'est le nombre de bénéficiaires pour chaque adhérent
`consult general`	C'est le nombre de consultations des généralistes correspondants à chaque adhèrent
`consult special`	C'est le nombre de consultations des spécialistes correspondant à chaque adhèrent
`depense`	Ce sont les dépenses correspondantes à chaque adhérent
`nbr prestation`	C'est le nombre de prestation pour chaque adhérent
`nbr bulletin soin`	C'est le nombre de bulletins de soin pour chaque adhérent
`nbr lignes pharma`	C'est le nombre de ligne pharmacie pour chaque adhérent
`nbr factures pharma`	C'est le nombre de factures pharmacie envoyées par un adhérent
`nbr bs labo`	C'est le nombre de bulletins de soin des laboratoires pour chaque adhérent
`hospit`	C'est le nombre de fois que l'adhérent ou ses ayant droit soient hospitalisés

TABLE 2.2 – Table des adhérents

✎ Ces variables sont toutes quantitatives

✎ Certains algorithmes d'apprentissage non-supervisé nécessitent des transformations des va-

riables quantitatives tel que la standardisation pour le K-means, d'autre n'exigent pas ces transformations, tel que l'Isolation Forest que nous allons utiliser pour la détection des aberrations.

- ✎ Les entrées de l'algorithme de détection d'anomalies vont être des ratios tels que :
  - ✏ `dépense/nbr malade`
  - ✏ `nbr bs/nbr malade`
  - ✏ `nbr ligne pharm/facture pharm`
  - ✏ `nbr prestation/nbr malade`
  - ✏ `depense/hospit`

Ces variables sont utilisées afin de prendre en considération le nombre d'ayant droit et le nombre de fois que l'adhérent et ses ayant droit soient hospitalisés.

2.2 Analyses descriptives

Dans cette section, nous allons effectuer une analyse descriptive des différentes variables. Cette étape est une étape cruciale à réaliser en amont de la mise en place du modèle de détection des comportements atypiques des assurés.

Distribution des dépenses des adhérents

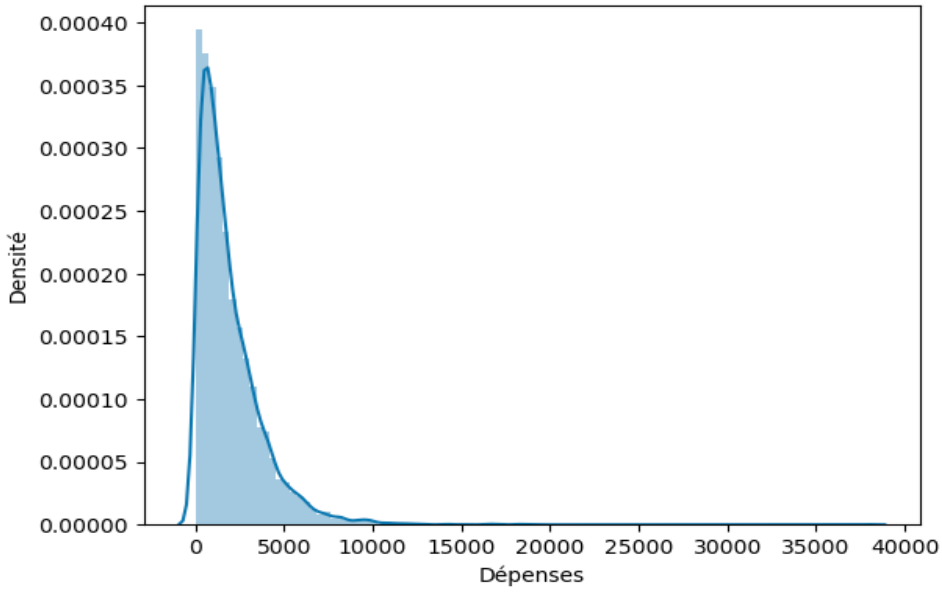


FIGURE 2.1 – Distribution des dépenses des assurés

La Figure 2.1 montre une distribution qui est semblable à une distribution Gamma, où la majorité des dépenses, par an, par les assurés et leurs ayant droit sont inférieure à 5 000 TND, en possédant des montants très élevé qui présentent les évènements aberrants.

Distribution des nombres de prestations des assurés

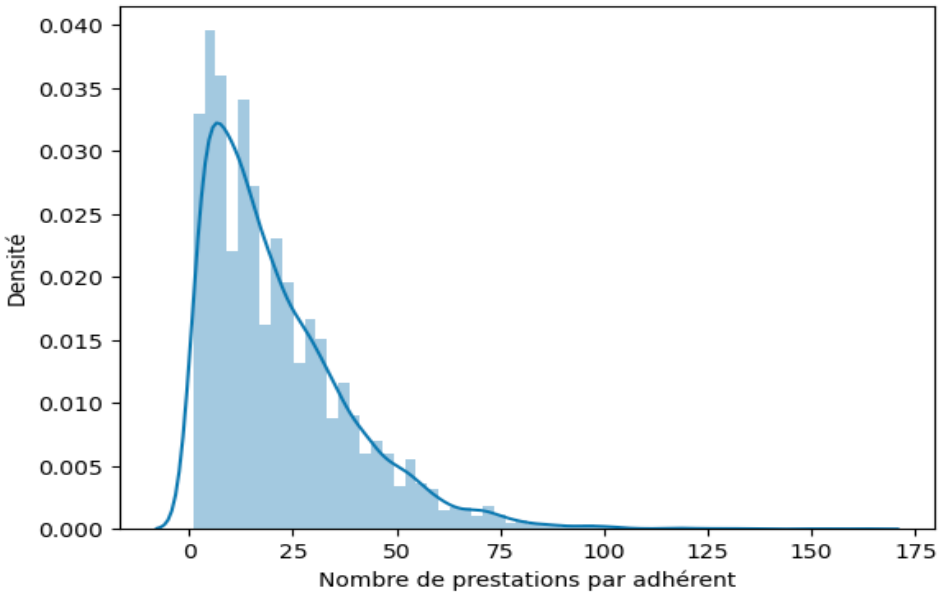


FIGURE 2.2 – Distribution de nombre de prestations par assurés

De même que pour les nombres de prestations par assurés, nous trouvons qu'ils obéissent à une loi de Poisson. La majorité des assurés ont un nombre de prestations, par an, inférieur à 50 en présence de quelques aberrations (Figure 2.2).

Distribution des âges des assurés

Nous remarquons une diversité dans la distribution des âges des adhérents. Cela nous montre de manière claire l'hétérogénéité de notre population (Figure 2.3).

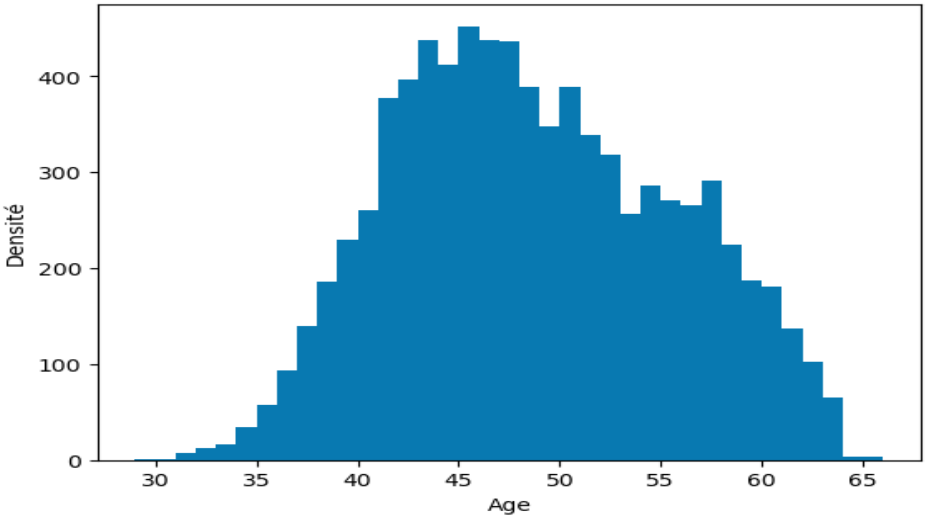


FIGURE 2.3 – Distribution des âges des adhérents

Distribution du nombre de factures pharmacie

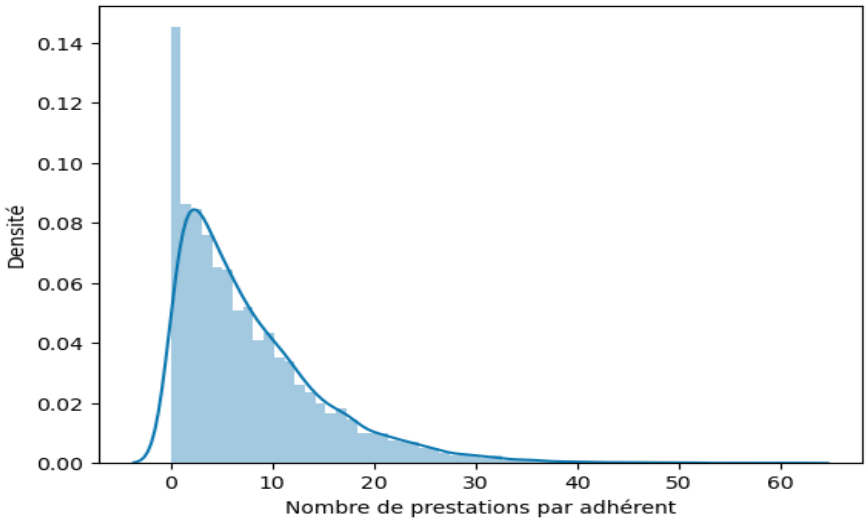


FIGURE 2.4 – Distribution du nombre de factures pharmacie par adhérents

La Figure 2.4 montre que la distribution du nombre de factures pharmacie par adhérent ressemble à une Poisson.

Distribution de nombre de consultations des généralistes par adhérent

Cette variable présente le nombre de fois qu'un adhérent et ses ayant droit consultent un généraliste.

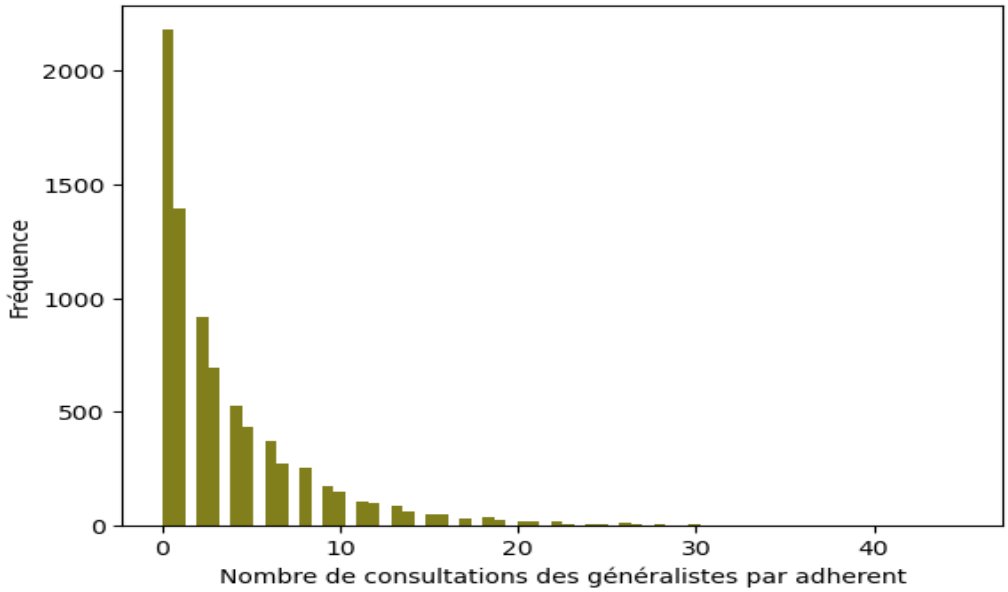


FIGURE 2.5 – Distribution du nombre de consultations des généralistes

La Figure 2.5 montre que le nombre de consultations des généralistes par adhérent suit une loi semblable à une Poisson, où les fréquences faibles sont les plus souvent présentes, alors que les fréquences élevées se présentent en tant qu'aberrations.

Fréquence des hospitalisations par adhérent

Cette variable décrit le nombre de fois qu'un assuré et ses ayant droit soient hospitalisés. Cette variable est très importante parce que, généralement, quelqu'un qui est hospitalisé n'est pas considéré comme fraudeur. En effet, il est normal qu'une personne hospitalisée, dépense beaucoup en matière de soins médicaux.

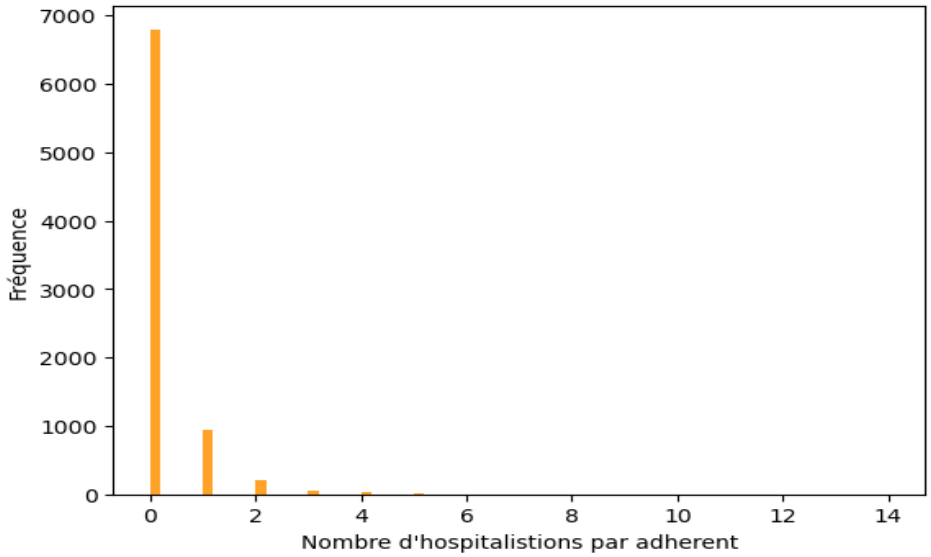


FIGURE 2.6 – Fréquence des hospitalisations par adhérent

Nous voyons sur la Figure 2.6 que la majorité des assurés n'étaient jamais hospitalisés, ceux sur lesquels nous allons nous intéresser.

Nombre de malades par adhérent

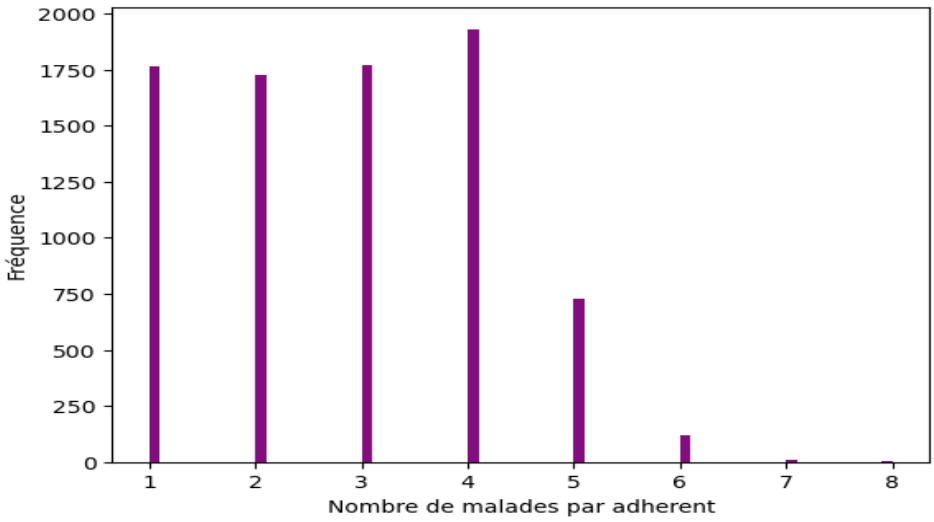


FIGURE 2.7 – Fréquences des nombres de malades par adhérent

Cette variable nous donne une indication sur le nombre de bénéficiaires par adhérent. Cette variable est importante dans l'étude des comportements atypiques, parce que c'est normal qu'une grande famille (en terme de membres) dépense plus qu'une petite famille (Figure 2.7).

Corrélations entre les variables

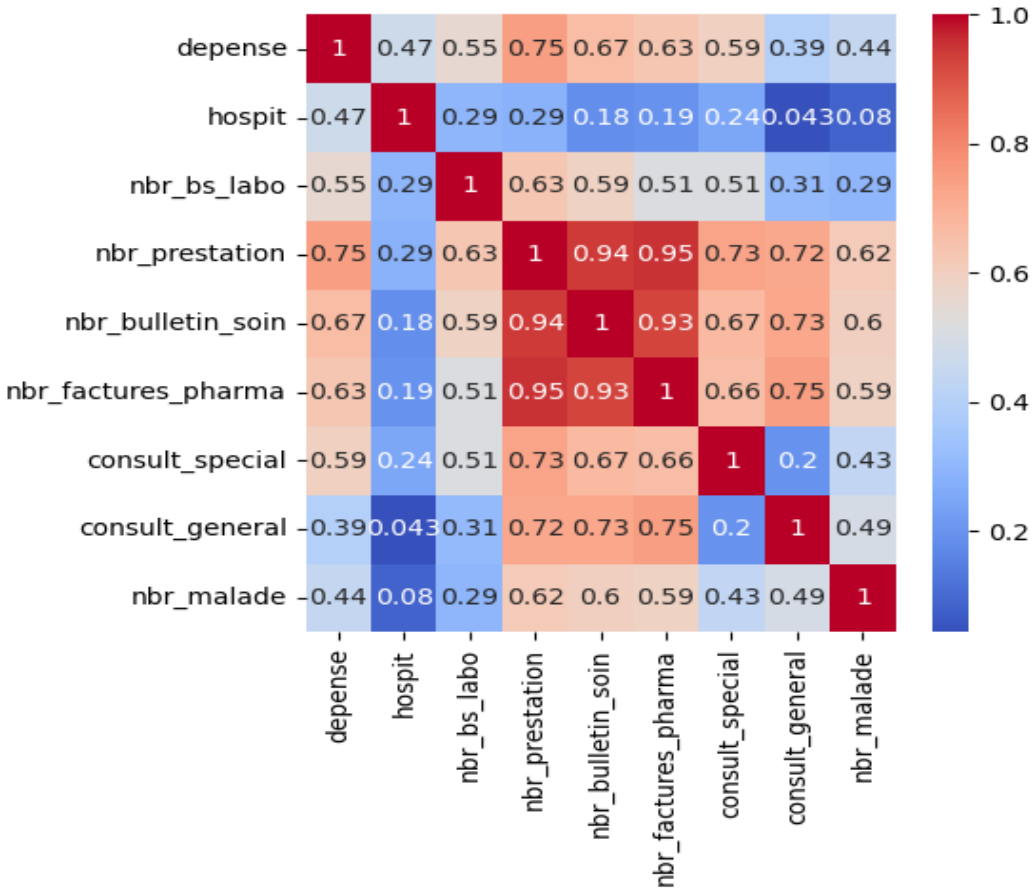


FIGURE 2.8 – Matrice des corrélations entre les variables

La Figure 2.8 représente la matrice des corrélations entre les différentes variables. Cette matrice montre des fortes corrélations entre les variables `nbr prestation, nbr bulletin soin, nbr factures pharma`. On remarque aussi que la variable `depense` est la seule variable qui est corrélée, légèrement, avec la variable `hospit`. Cela peut-être interprété par le nombre d'hospitalisations qui est généralement faible, ceci est claire dans la Figure 2.6. Cette matrice permet, dans des cas, la séléction des variables à implémenter dans les modèles (*Feature selection*). Sauf que dans notre cas, nous allons nous baser sur des connaissances métier pour le choix des variables.

Deuxième partie

Calcul des provisions pour sinistres à payer (PSAP)

Chapitre 3

Modèles de provisionnement

Un des rôles majeurs de l'actuaire dans le domaine de l'assurance non vie est le calcul des provisions, et plus spécifiquement, les provisions pour sinistres à payer (PSAP). Du point de vue de la solvabilité, la compagnie d'assurance se doit de réserver le plus possible de provision, mais en terme de performance et de rentabilité vis-à-vis des actionnaires, elle vise à réserver un montant minimum de provisions. Une bonne estimation de ces provisions est donc un enjeu pour l'entreprise.

Actuellement, les techniques utilisées pour estimer ces provisions reposent essentiellement sur des approches déterministes basées sur les triangles de liquidation. Ainsi, la méthode Chain Ladder est la méthode la plus répandue dans le secteur de l'assurance. Cette méthode est simple à mettre en oeuvre et est facile à interpréter. Cependant, elle repose sur des hypothèses d'homogénéité très fortes, ce qui peut entraîner des lacunes méthodologiques dans divers scénarios [19]. L'objectif dans ce chapitre est donc de proposer une approche actuarielle alternative, adaptée aux sinistres hétérogènes, qui a pour but de renforcer la méthode standard de Chain Ladder.

3.1 Triangles de liquidation

Les méthodes de provisionnement sont toutes basées sur la sinistralité passée de la branche étudiée. Ces données sont organisées dans un format triangulaire connu sous le nom de triangle de liquidation. Nous utilisons les notations :

- ✓ i l'indice du mois de survenance, $i = 1..n$
- ✓ j l'indice du mois de développement, $j = 1..n$

- ✓ $Y_{i,j}$ le montant des sinistres survenus le mois i et payés au bout de j mois.
- ✓ $C_{i,j}$ le montant cumulé des sinistres survenus le mois i et payés durant les mois $1..j$. On aura $C_{i,j} = Y_{i,1} + ... + Y_{i,j}$.

Le triangle de liquidation, qu'il soit cumulé ou non cumulé, se présente comme dans la Figure 3.1 :

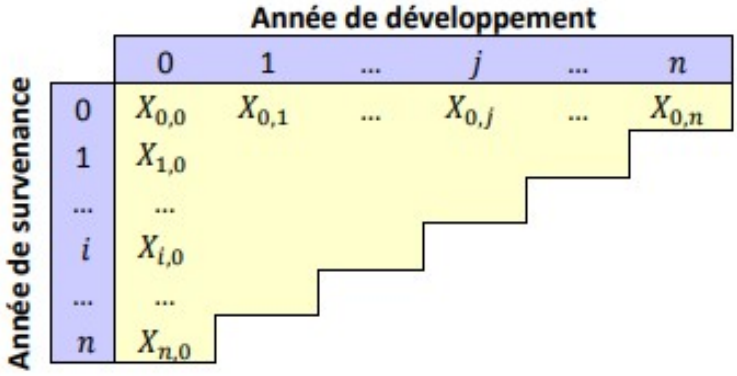


FIGURE 3.1 – Structure du triangle de liquidation

Le processus de détermination des provisions consiste à prévoir la valeur finale des sinistres et à estimer les paiements restant à effectuer. Pour ce faire, on suppose que les sinistres survenus au cours d'un mois donné sont entièrement réglés au cours des n mois suivants. L'objectif est donc de remplir la partie inférieure du triangle. On cherche donc $\widehat{X_{i,j}}$ pour $i + j > n$.

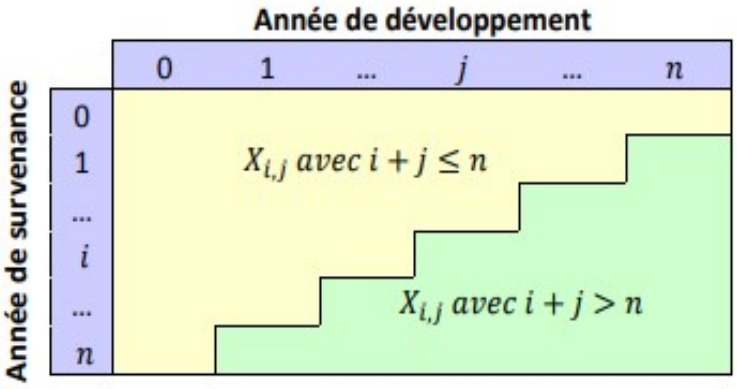


FIGURE 3.2 – Structure du triangle complété de liquidation

où $\widehat{X}_{i,j} = \begin{cases} \widehat{Y}_{i,j} \text{ dans le cas d'un triangle décumulé} \\ \widehat{C}_{i,j} \text{ dans le cas d'un triangle cumulé} \end{cases}$

Après avoir estimer ces paiements futurs, on calcule le montant de provision à effectuer pour le mois de survenance i :

$$\begin{aligned} \widehat{R}_i &= \widehat{C}_{i,n} - C_{i,n-i} && (3.1) \\ &= \sum_{j=n-i}^{n} \widehat{Y}_{i,j} && (3.2) \end{aligned}$$

Le montant total nécessaire de réserves, est égal donc à :

$$\hat{R} = \sum_{i=1}^{n} \widehat{R}_i \qquad (3.3)$$

3.2 Méthode déterministe Chain Ladder

Les modèles déterministes considèrent uniquement les valeurs des variables à expliquer, sans tenir compte du caractère aléatoire qui leur est associé. Ces modèles partent de l'hypothèse que le délai entre la survenance d'un sinistre et son règlement reste constant, quelle que soit l'année considérée.

3.2.1 Description de la méthode Chain Ladder

Cette méthode est la plus couramment utilisée par les compagnies d'assurance en raison de sa simplicité de mise en œuvre [7]. Le concept fondamental de cette approche repose sur l'utilisation des facteurs de développement, en anglais appelés *link-ratios*.

Les hypothèses de base de cette méthode sont :

(H1) : Les mois de survenance sont indépendantes entre elles. Cette hypothèse est considéré toujours valide, en pratique.

(H2) : Les mois de développement sont des variables explicatives du comportement des sinistres futurs.

La méthode Chain Ladder suppose que les ratios $\dfrac{C_{i,j+1}}{C_{i,j}}$ sont indépendants de l'année de survenance i. C'est-à-dire que :

$$\frac{C_{0,j+1}}{C_{0,j}} = \frac{C_{1,j+1}}{C_{1,j}} = ... = \frac{C_{n-j-1,j+1}}{C_{n-j-1,j}}$$

Les facteurs de développement peuvent être estimés à l'aide des observations par la formule suivante :

$$\widehat{f}_j = \sum_{i=0}^{n-j-1} \frac{C_{i,j+1}}{C_{i,j}} \qquad pour\ j = 0, .., n-1 \tag{3.4}$$

En utilisant les facteurs de développement précédemment estimés, il devient possible d'obtenir une estimation des montants futurs, pour $i + j > n$:

$$\widehat{C}_{i,j} = C_{i,n-i}(\widehat{f}_{n-i} \times ... \times \widehat{f}_{n-1}) \qquad \text{pour } i + j > n.$$

Les réserves (provisions) à effectuer chaque mois i et le montant total des réserves, équivalent à la PSAP sont déterminés comme suit :

$$\widehat{R}_i = \widehat{C}_{i,n} - C_{i,n-i} \qquad \text{et} \qquad \widehat{R} = \textstyle\sum_{i=1}^{n} \widehat{R}_i.$$

3.2.2 Validation des hypothèses (d-triangle)

La méthode Chain Ladder ne peut-être appliquée que si les hypothèses mentionnées précédemment soient vérifiées.

Pour vérifier l'hypothèse (H2), on peut utiliser le triangle des facteurs individuels (d-triangle). Le d-triangle comprend les facteurs individuels suivants :

$$f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} \qquad \text{pour } i + j \leq n - 1$$

Une fois le triangle calculé, on peut dire que l'hypothèse d'indépendance est vérifiée si, pour $j = 0, ..., n-2$, les élément de la j^{eme} colonne du d-triangle sont *sensiblement* constants.

3.2.3 Critique de la méthode

Bien que cette méthode soit très répandue dans le monde de l'assurance en raison de sa facilité de mise en œuvre, elle présente néanmoins plusieurs faiblesses inhérentes.

- ✗ L'hypothèse de l'indépendance des dates de survenance des sinistres est une hypothèse très forte. En effet, les conditions suivantes doivent être toutes présentes pour qu'elle soit valide [14].
  - ☛ L'existence d'une régularité suffisante dans le passé. Par exemple, il ne doit pas y avoir un changement important dans la gestion des sinistres.
  - ☛ La branche ne doit pas être volatile : il est difficile de traiter les sinistres graves, en particulier s'il sont ponctuels, par cette méthode.

☛ Les données du portefeuille doivent être nombreuses et fiables.

✗ L'incertitude de l'estimation augmente avec le nombre de mois de survenance. Plus précisément, pour le mois le plus récent, le coefficient multiplicatif est le résultat de la multiplication de $(n-1)$ estimations des facteurs de développement. Cette incertitude devient plus prononcée lorsqu'il s'agit de branches caractérisées par des périodes de développement prolongées.

La méthode Chain Ladder est largement reconnue comme la méthode déterministe la plus simple pour calculer les réserves et est couramment utilisée par les professionnels de l'assurance. Cependant, plusieurs autres méthodes déterministes, telles que les modèles de Bornhuetter-Ferguson, de coût moyen et de De Vylder, ont également été développées. Ces méthodes déterministes, y compris la méthode Chain Ladder, ont pour limite de ne prendre en compte que les valeurs des variables à expliquer et d'ignorer leur nature aléatoire. Par conséquent, ces méthodes fournissent des estimations des provisions moyennes sans tenir compte de la volatilité de ces estimations.
Cette limite a conduit à l'émergence de modèles stochastiques, qui considèrent les composantes du triangle comme de véritables variables aléatoires. Ces modèles stochastiques supposent que ces variables suivent une distribution de probabilité spécifique déterminée par les données observées du triangle. Parmi les modèles stochastiques les plus connus, on peut citer le modèle de Mack, le modèle log-normal et les modèles linéaires généralisés.
Dans la suite de ce chapitre, nous nous concentrerons sur le modèle de Mack et les GLMs.

3.3 Mack Chain Ladder

Ce modèle proposé par Mack en 1993 est une approche non paramétrique qui permet d'estimer la marge d'erreur associée aux montants des réserves. Ce modèle est communément appelé distribution *Free Chain-Ladder*, car il n'impose aucune hypothèse de distribution spécifique sur les composantes du triangle.

3.3.1 Hypothèses

Le modèle de Mack permet d'estimer les erreurs commises lors de l'evaluation des provisions. Pour cela, il pose trois hypothèses :

(H1) : Indépendance des dates de survenance des sinistres :

Pour $i_1 \neq i_2$, les v.a $(C_{i_1,j})_{j=1..n}$ et $(C_{i_2,j})_{j=1..n}$ sont indépendantes.

(H2) : L'espérance conditionnelle de $C_{i,j+1}$ sachant $C_{i,1}, ..., C_{i,j}$, est liée à la dernière observation $C_{i,j}$ par un facteur μ_j par la relation :

$$\mathbb{E}[C_{i,j+1}|C_{i,1}, ..., C_{i,j}] = \mu_j \times C_{i,j}$$

(H3) : La variance de $C_{i,j+1}$ sachant $C_{i,1}, ..., C_{i,j}$, est liée à la dernière observation $C_{i,j}$ par un facteur σ_j par la relation :

$$\mathbb{V}(C_{i,j+1}|C_{i,1}, ..., C_{i,j}) = \sigma_j^2 \times C_{i,j}.$$

3.3.2 Principe

On prend pour estimateurs des paramètres du modèle stochastique les même paramètres que pour la méthode Chain Ladder standard :

$$\widehat{f_j} = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}} \qquad pour j = 0, .., n-1 \tag{3.5}$$

On obtient exactement les même estimations des réserves qu'avec la méthode de Chain Ladder classique :

$$\mathbb{E}[C_{i,j+1}|C_{i,1}, ..., C_{i,j}] = \widehat{C}_{i,n} = C_{i,n-i} \times \widehat{f}_{n-i} \times ... \times \widehat{f}_{n-1} \text{ pour } i = 1, ..., n.$$

$$\widehat{R}_i = \widehat{C}_{i,n} - C_{i,n-i} \tag{3.6}$$

$$\widehat{R} = \sum_{i=1}^{n} \widehat{R}_i \tag{3.7}$$

3.3.3 Erreur de prévision

En se basant sur les estimations effectuées et l'hypothèse (H3), il est possible d'étudier l'erreur de prévision en calculant la distance moyenne entre l'estimateur $\widehat{R}_i$ et la vraie valeur R_i.

On utilise $\widehat{\sigma}_j^2$ comme estimateur de $\widehat{\sigma}_j$ pour $j = 0, ..., n-2$:

$$\widehat{\sigma}_j^2 = \frac{1}{n-j-1} \sum_{i=0}^{n-j-1} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - \widehat{f}_j \right)^2 \tag{3.8}$$

$$\widehat{\sigma}_{n-1}^2 = min \left(\frac{\widehat{\sigma}_{n-2}^4}{\widehat{\sigma}_{n-3}^2}, \widehat{\sigma}_{n-2}^2, \widehat{\sigma}_{n-3}^2 \right) \tag{3.9}$$

On définit l'erreur quadratique moyenne (MSE [1]), pour $i = 1, ..., n$ et $j = 0, ..., n-1$, du montant des réserves pour l'année i par la formule :

$$MSE(\widehat{R}_i) = \mathbb{E}[(\widehat{R}_i - R_i)^2 | D] \quad avec \quad D = \{C_{i,j} | i + j \leq n\}. \tag{3.10}$$

L'estimation de cette erreur est :

$$\widehat{MSE}(\widehat{R}_i) = C_{i,n}^2 \sum_{j=n-i}^{n-1} \frac{\widehat{\sigma}_j^2}{\widehat{f}_j^2} \left(\frac{1}{\widehat{C}_{i,j}} + \frac{1}{\sum_{k=0}^{n-j-1} C_{k,j}} \right) \tag{3.11}$$

L'écart type pour chaque variable $\widehat{R}_i$, SE [2] est estimée par :

$$\widehat{SE}(\widehat{R}_i) = \sqrt{\widehat{MSE}(\widehat{R}_i)} \tag{3.12}$$

Ainsi, l'estimation de l'erreur quadratique moyenne des réserves totales :

$$\widehat{MSE}(\widehat{R}) = \sum_{i=1}^{n} \left(\widehat{MSE}(\widehat{R}_i) + \widehat{C}_{i,n} \left(\sum_{k=i+1}^{n} \widehat{C}_{k,n} \right) \times \sum_{j=n-i}^{n-1} \frac{2\widehat{\sigma}_j^2 / \widehat{f}_j^2}{\sum_{k=0}^{n-j-1} C_{k,j}} \right) \tag{3.13}$$

Ainsi, l'estimation de l'erreur standard est :

$$\widehat{SE}(\widehat{R}) = \sqrt{\widehat{MSE}(\widehat{R})} \tag{3.14}$$

Cette métrique représente l'erreur dans l'estimation de la réserve totale. L'objectif est donc de minimiser cet indicateur. Une valeur élevée de cette métrique indique que le modèle est moins adapté à l'évaluation des provisions du triangle.

3.3.4 Intervalle de confiance

Thomas Mack propose de construire des intervalles de prédiction pour la provision en faisant des hypothèses sur la distribution prédictive conditionnelle. Cependant, malgré la nature non paramétrique de ce modèle, la sélection d'une distribution spécifique reste très arbitraire.

Les deux paramètres de la distribution sont les estimations de la moyenne et de l'écart type conditionnels de R,

$$\widehat{\mathbb{E}}[R] = \widehat{R} \qquad \text{et} \qquad \widehat{SE}(\widehat{R}).$$

1. Mean Squared Error
2. Standard Error

Par exemple, si on choisit une distribution normale $\mathcal{N}(\mu, \sigma^2)$, où $\mu = \widehat{R}$ et $\sigma = \widehat{SE}(\widehat{R})$, alors l'intervalle de confiance à un seuil de 95% pour R est :

$$\left[\widehat{R} - 1,96 \times \widehat{SE}(\widehat{R}) \; ; \; \widehat{R} + 1,96 \times \widehat{SE}(\widehat{R})\right]$$

L'estimation du quartile d'ordre α de R se fait par :

$$\widehat{q}_\alpha(R) = \widehat{R} + \widehat{SE}(\widehat{R}) q_\alpha \qquad \text{où } q_\alpha \text{ est le quantile de la loi Normale standard.}$$

Si on choisit une lognormale Log$\mathcal{N}(\mu, \sigma^2)$, cela veut dire que Ln(R) suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. Alors μ et σ^2 sont déterminés par le système :

$$\begin{cases} e^{\mu + \frac{\sigma^2}{2}} = \widehat{R} \\ e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) = (\widehat{SE}(\widehat{R}))^2 \end{cases}.$$

D'où :

$$\sigma^2 = ln\left(1 + \frac{(\widehat{SE}(\widehat{R}))^2}{\widehat{R}^2}\right) \text{ et } \mu = ln(\widehat{R}) - \frac{\sigma^2}{2}.$$

L'intervalle de confiance pour R est alors egal à :

$$\left[e^{\mu - 1,96 \times \sigma} \; ; \; e^{\mu + 1,96 \times \sigma}\right]$$

Et le quartile d'ordre α de R est : $e^{\mu + q_\alpha . \sigma}$.

3.3.5 Validation des hypothèses

Dans la partie consacrée à l'utilisation de la méthode Chain Ladder, nous avons précédemment examiné la procédure de validation de l'hypothèse (H1) d'indépendance entre les années de survenance.

Pour ce qui concerne la validation de l'hypothèse (H2), qui suppose une régression linéaire entre les montants cumulés, pour tout j, on vérifie que les couples $(C_{i,j}, C_{i,j+1})_{i=1,...,n-j-1}$ sont *sensiblement* alignés sur une droite passante par l'origine.

Il reste à vérifier l'hypothèse (H3). Pour se faire, on trace le graphique des résidus normalisés en fonction de $C_{i,j}$, pour tout j fixé :

$$r_{i,j} = \frac{(C_{i,j+1} - \widehat{f}_j C_{i,j})}{\sqrt{C_{i,j}}} \quad \text{pour } i = 1, ..., n - j - 1 \tag{3.15}$$

Pour que l'hypothèse ne soit pas rejetée, il est nécessaire que les résidus ne démontrent aucune structure non aléatoire.

3.4 Modèles Linéaires Généralisés (GLM)

Les modèles linéaires généralisés (GLM) sont des techniques statistiques qui étendent le concept de régression linéaire.

Dans un premier temps, nous allons présenter le modèle linéaire standard. Ensuite, nous détaillerons le modèle linéaire généralisé.

3.4.1 Modèle linéaire classique

Ce modèle a pour but d'expliquer la variable endogène (à expliquer) Y à l'aide d'une ou plusieurs variables exogènes (explicatives), selon la formule suivante :

$$Y = X \times \beta + \epsilon$$

où :

- Y est un vecteur de $\mathbb{R}^n$
- X est une matrice de $\mathbb{R}^n \times \mathbb{R}^{p+1}$, contenant les p variables prédicteurs
- β est le vecteur des paramètres à estimer
- ϵ est un terme d'erreur, contient les résidus.

On a :

$$\beta = (\beta_0, \ldots, \beta_p)$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & a_{11} & \cdots & a_{1p} \\ 1 & a_{21} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_{n1} & \cdots & a_{np} \end{pmatrix}$$

Par la méthode des moindres carrés, on obtient un estimateur de β :

$$\widehat{\beta} = (X^T X)^{-1} X^T Y.$$

On a alors $\widehat{\epsilon} = Y - \widehat{Y}$ le vecteur des n résidus.

Le coefficient de détermination R^2 permet de quantifier la qualité de l'ajustement linéaire.

$$R^2 = \frac{\sum \widehat{y}_i^2}{\sum y_i^2}$$

où SCE est la somme des carrés expliqués $\sum \widehat{y}_i^2$ et SCT est la somme des carrés totaux $\sum y_i^2$.

Plus la valeur de R^2 est proche de 1, plus l'ajustement est bon.

3.4.2 Modèles Linéaires Généralisés

Les Modèles Linéaires Généralisés ou *Generalized Linear Models* (GLM) étendent le modèle linéaire gaussien en permettant l'utilisation d'autres lois (conditionnelles) que la loi gaussienne. Ces modèles intègrent trois éléments clés : la composante aléatoire, la composante systématique et la fonction lien.

La composante aléatoire

Cette composante correspond à la variable à expliquer Y. On note $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ où les densités de Y_i appartiennent à une famille de lois spécifiques à un GLM.
Pour ce qui rapporte notre travail, nous aurons à expliquer les montants des sinistralités mensuelles selon leurs mois de paiement.

La composante systématique

Les variables $X_1, ..., X_p$ sont des variables explicatives pour le modèle. Pour chaque variable, on a n observations. soit $x = (x_1, ..., x_n)$ une observation de ces variables. Le prédicteur linéaire associé à cette observation sera défini par :

$$\nu(x) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i$$

où $\beta = (\beta_0, ..., \beta_p)$ est un vecteur de paramètres à estimés.
Dans notre cas, provisionnement à deux variables exogènes qualitatives, la composante systématique sera :

$$\nu_{i,j} = \mu + \alpha_i + \beta_j \qquad i, j = 0, ..., n.$$

On obtient alors le modèle saturé.

La fonction lien

Cette fonction, noté g, est déterministe et est strictement monotone définie sur $\mathbb{R}$. Cette fonction fait le lien entre la composante déterministe et la composante systématique. Elle relie l'espérance de la variable à expliquer Y, noté μ, au prédicteur linéaire ν, telle que :

$$\nu(x) = g(\mu).$$

La fonction lien vérifiant : $\mu = g^{-1}(\theta) \iff \theta = \nu(X)$ est dite fonction de lien canonique. dans La Table 3.1 nous présentons les fonctions de lien canoniques des différentes distribution :

Loi de probabilité	$V(\mu)$	**Fonction lien canonique**
Poisson	μ	$ln(\mu)$
Normale	1	μ
Gamma	μ^2	$\frac{1}{\mu}$
Inverse Gaussienne	μ^3	$\frac{1}{\mu^2}$
Binomiale	$\mu(1-\mu)$	$ln(\frac{\mu}{1-\mu})$

TABLE 3.1 – Fonctions de lien canoniques

Famille de lois

Les lois possibles d'un modèle GLM font partie de la famille exponentielle, qui a été introduite par Nelder et Wedderburn en 1972.

Une loi de probabilité appartient à la famille exponentielle si sa fonction densité s'écrit ainsi :

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

où

- $\theta \in \mathbb{R}$ est dit paramètre canonique ou paramètre naturel.
- $\phi \in \mathbb{R}_+^*$ est dit paramètre de dispersion.
- a, b, c des fonctions connues, dérivables, où b est une fonction 3 fois dérivable et sa dérivé première est inversible.

Pour Y de densité $f_{\theta,\phi}$ de forme exponentielle, telle que définie ci-dessus, on obtient :

$$\begin{aligned}
\mathbb{E}(Y) &= b^{'}(\theta) && (3.16)\\
&= \mu && (3.17)\\
\mathbb{V}(Y) &= b^{''}(\theta)\, a(\phi) && (3.18)\\
&= \mathbb{V}(\mu) && (3.19)
\end{aligned}$$

Estimation par maximum de vraisemblance

Cette méthode est utilisée pour estimer les paramètres $\beta_0, ..., \beta_p$.
La fonction de log-vraisemblance est définie comme suit :

$$l(y, \theta, \phi) = ln(f_{\theta,\phi}(y))$$

Dans le cadre des GLM, Y est telle que sa densité de probabilité est définie précédemment. On aura donc :

$$l(y, \theta, \phi) = \frac{(y\theta - b(\theta))}{a(\phi)} + c(y, \phi)$$

On cherche à déterminer les $\widehat{\beta}_i$ qui maximisent la log-vraisemblance. On remarque que ces $\widehat{\beta}_i$ n'apparaissent pas explicitement dans l'expression de la log-vraisemblance, on va utiliser alors les dérivées partielles. Pour cela, on décompose en dérivées partielles :

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \nu_i} \times \frac{\partial \nu_i}{\partial \beta_j}. \tag{3.20}$$

Problème rencontré dans notre cas : Triangle de provisionnement

Un des inconvénients rencontré est la quantité de paramètres. Dans le cas des triangles de provisionnement, il y en a $2n + 1$ paramètres à estimer, où n correspond au nombre de modalités pour chaque variable exogène (mois de survenance et de développement). Ce qui peut augmenter le risque d'erreur de prédiction. Si on a trois modalités pour chaque variable, l'équation $Y = X\beta + \epsilon$ aurait pour solution : $\widehat{\beta} = (\widehat{\mu}, \widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)$, où :

- ☞ $\widehat{\mu}$ est la moyenne des incréments.
- ☞ $\widehat{\alpha}_i$ est le coefficient qui correspond au mois de survenance i.
- ☞ $\widehat{\beta}_j$ est le coefficient qui correspond au mois de développement j.

Qualité de la régression

Pour évaluer la précision d'une régression, nous définissons une fonction de risque R qui quantifie l'écart entre la valeur observée Y et la valeur estimée $\widehat{Y} = \widehat{\mu}$. Généralement, nous utilisons la norme L^2 qui correspond à l'erreur quadratique $R(Y, \widehat{Y}) = (Y - \widehat{Y})^2$.
Nous utilisons ces types de résidus :

- ✗ Résidus de déviation : $\widehat{\epsilon}_i = (Y_i - \widehat{Y}_i)$

✗ Résidus de Pearson : $\widehat{\epsilon}_i = \dfrac{Y_i - \widehat{Y}_i}{\sqrt{V(Y_i)}}$

✗ Résidus de déviance : $\widehat{\epsilon}_i = signe(Y_i - \widehat{Y}_i)\sqrt{d_i}$

Nous allons utiliser le critère d'information d'Akaike (AIC) pour pouvoir choisir le meilleur modèle.

$$AIC = -2log(L) + 2k$$

Où k et le nombre de paramètres.

Le meilleur modèle est celui qui l'AIC le plus faible.

Bootstrapping

Dans cette partie, nous proposons de construire des intervalles des confiance pour les provisions. Pour cela, nous allons exploiter les résidus afin de simuler des faux triangles de liquidation. Nous considérons les erreurs de déviation $\widehat{\epsilon}_i = (Y_i - \widehat{Y}_i)$. L'échantillonnage de l'ensemble des résidus se fait à plusieurs reprises mais avec remise. Soit :

$$Y_i = \widehat{Y}_i + \widehat{\epsilon}_l$$

Finalement nous nous retrouvons avec une distribution des PSAP.

3.5 Conclusion

Ce chapitre a introduit les fondements mathématiques derrière les différents modèles de provisionnement dont les résultats vont être discutés dans le chapitre suivant. En effet, ce sont les modèles Chain Ladder et GLM qui ont été présenté en détails. La compréhension des fondements mathématiques permet l'amélioration, la validation et le déploiement.

Chapitre 4

Application et validation

4.1 Introduction

Dans ce chapitre, nous allons présenter les résultats, les analyses et les conclusions sur les performances des modèles de provisionnement dont les fondements mathématiques sont détaillés dans le chapitre précédent. Nous allons, essentiellement, analyser les résultats du modèle GLM et les comparer au modèle de Mack appliqué sur les même données.

Les données utilisées pour la validation sont relatives aux sinistres survenus en 2021.

	1	2	3	4	5	6	7	8	9	10	11	12
2021-01	22,192	491,545	589,726	606,042	609,975	611,715	611,798	617,524	619,421	619,733	619,939	620,069
2021-02	118,352	529,635	655,958	661,016	661,108	661,320	661,338	661,376	661,376	661,376	661,829	
2021-03	70,095	692,066	749,190	751,706	753,391	753,806	753,912	754,549	754,549	754,549		
2021-04	221,281	508,016	548,317	553,345	553,617	555,046	555,300	555,665	555,714			
2021-05	141,724	496,427	540,424	546,579	548,583	549,251	549,373	549,373				
2021-06	197,953	535,012	593,549	596,310	606,753	606,822	609,804					
2021-07	113,865	378,193	437,649	440,823	441,842	442,213						
2021-08	118,163	472,047	520,523	528,531	530,082							
2021-09	234,209	573,935	668,905	678,864								
2021-10	93,020	531,630	630,724									
2021-11	126,949	587,295										
2021-12	91,318											

FIGURE 4.1 – Triangle de test

Les performances sont évaluées en comparant les résultats des modèles au montant total payé en 2022 relatif aux sinistres survenus en 2021.

4.2 Méthode de Chain Ladder

En ce qui concerne cette méthode, on va se contenter de la validation des hypothèses (H1) et (H2). Les résultats seront discutés dans la section suivante car cette méthode et le modèle de Mack donnent exactement les même estimations comme nous l'avons mentionné précédemment. En pratique, l'hypothèse d'indépendance (H1) est toujours valide.

d-triangle

	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12
2021-01	22.1494	1.1997	1.0277	1.0065	1.0029	1.0001	1.0094	1.0031	1.0005	1.0003	1.0002
2021-02	4.4751	1.2385	1.0077	1.0001	1.0003	1.0000	1.0001	1.0000	1.0000	1.0007	
2021-03	9.8732	1.0825	1.0034	1.0022	1.0006	1.0001	1.0008	1.0000	1.0000		
2021-04	2.2958	1.0793	1.0092	1.0005	1.0026	1.0005	1.0007	1.0001			
2021-05	3.5028	1.0886	1.0114	1.0037	1.0012	1.0002	1.0000				
2021-06	2.7027	1.1094	1.0047	1.0175	1.0001	1.0049					
2021-07	3.3214	1.1572	1.0073	1.0023	1.0008						
2021-08	3.9949	1.1027	1.0154	1.0029							
2021-09	2.4505	1.1655	1.0149								
2021-10	5.7152	1.1864									
2021-11	4.6262										

FIGURE 4.2 – Présentation du triangle des facteurs individuels : d-triangle

La validation de l'hypothèse (H2) sera faite à l'aide du d-triangle.

Le triangle des facteurs individuels montre que les éléments de la première colonne sont «non-sensiblement constant». Pour le reste des colonnes, les facteurs individuels seront de plus en plus stables. Cela montre la non-validité de l'hypothèse pour la transition du 1^{er} au 2^{eme} mois de développement, alors que pour le reste, l'hypothèse est valide.

Globalement, on peut dire que l'hypothèse d'indépendance est à rejetée. On va donc voir l'impact de cette hypothèse sur le modèle.

4.3 Modèle de Mack

Dans cette section, on va discuter les résultats du modèle Mack's Chain Ladder selon la validité de ses hypothèses (H1) et (H2).

4.3.1 Discution des hypothèses

L'hypothèse d'indépendance (H1) a été déjà discutée dans la section consacrée au modèle classique de Chain Ladder. En ce qui concerne l'hypothèse (H2), on va considérer les graphiques suivant qui représentent les corrélations entre les valeurs observées dans le mois de développement j et les valeurs du mois $j+1$.

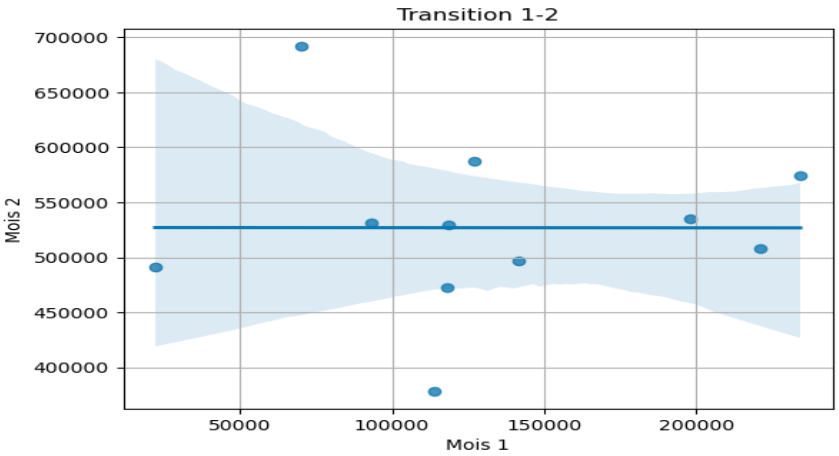


FIGURE 4.3 – Transition entre mois(1)-mois(2)

Nous voyons, dans la Figure 4.3, qu'il n'y a aucune corrélation entre les montants payés dans le premier mois de développement et ceux payés dans le deuxième mois de développement.

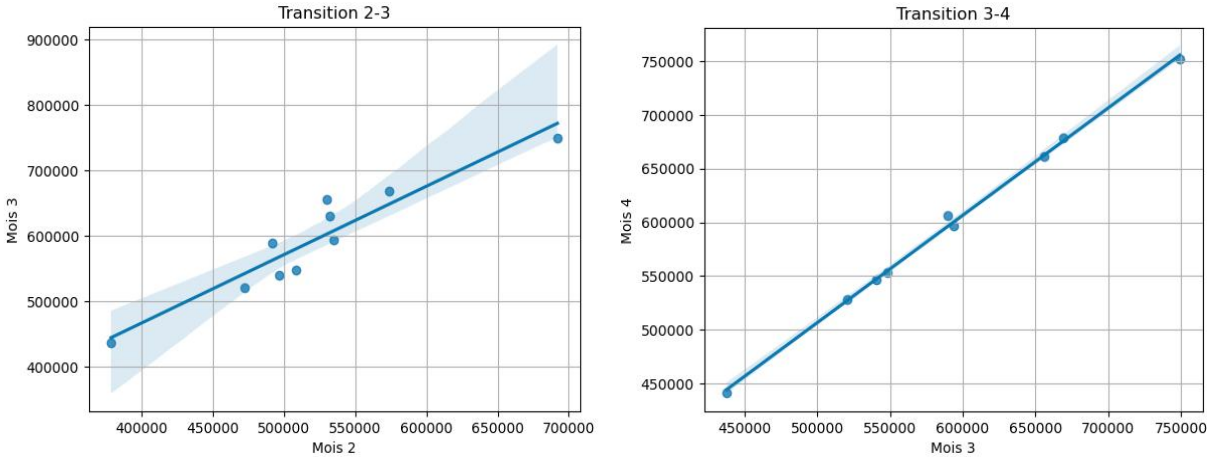


FIGURE 4.4 – Transition entre mois(2)-mois(3) et mois(3)-mois(4)

Les corrélations deviennent de plus en plus fortes en passant d'un mois de développement à un autre. On constate que les couples $(C_{i,j}, C_{i,j+1})_{j=2,3}$ sont proches d'une droite. D'où la validité de (H2) pour $j \in \{2,3\}$ i.e. le deuxième et le troisième mois.

Globalement, on va considérer l'hypothèse (H2) comme étant invalide. En effet, pour le mois le plus récent, le coefficient multiplicatif est le résultat de la multiplication de 11 estimations des facteurs de développement. L'estimation des IBNRs sera donc impactée par la transition du premier mois au deuxième mois de développement.

4.3.2 Application du modèle

Le modèle de Mack repose sur des triangles des données cumulées. On va donc utiliser le triangle présenté dans l'introduction de ce chapitre (Figure 4.1). L'estimation des facteurs de développement est donnée comme suit :

	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12
$\widehat{f}_j$	5.918	1.141	1.011	1.004	1.001	1.001	1.002	1.0008	1.0002	1.0005	1.0002

TABLE 4.1 – Coefficients de transition

La sinistralité survenue en un mois j est payée en grande partie au bout des quatre premiers mois. Après cela, les montants cumulés relatifs au mois j seront presque constants.

Estimation des montants ultimes

déja payé	Observé	Mack
620068.8376	620068.8376	620068.837600
661829.2450	661829.2450	661967.796560
754548.9144	755111.9524	755095.171603
555714.4083	555752.8923	556201.966486
549372.8436	551113.1376	550268.371474
609804.1090	611541.2730	612117.673273
442212.9070	445391.6300	444314.062414
530082.4290	532175.8000	533226.433715
678863.8730	682380.6630	685952.068820
630724.3810	652288.9550	644395.539205
587294.5080	720694.7470	683712.883461
91317.6880	701772.5030	422656.461209

FIGURE 4.5 – Résultats des estimation par Mack Chain Ladder

Les estimations des montants ultimes par le modèle de Mack Chain Ladder ont donné les résultats de la Figure 4.5. Le modèle de Mack Chain Ladder estime que le montant total de la sinistralité

relative à l'année 2021 est de 7 169 977 TND alors que la vraie valeur est de 7 490 122 TND, soit -320 145 TND qui correspond à l'écart. Ainsi pour le montant de provision, on retrouve le même écart puisque on a : IBNR = $Ultime - Dernier$.

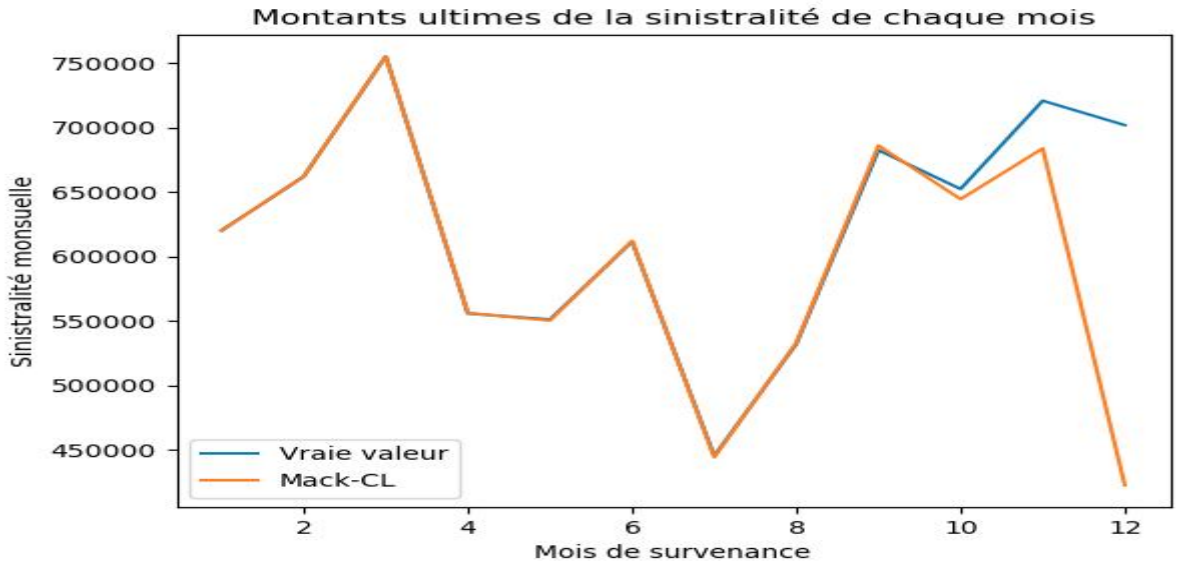


FIGURE 4.6 – Montants mensuels ultimes estimés par le modèle de Mack

Le schéma ci-dessus montre que l'écart entre le montant estimé et le montant observé réellement augmente pour les mois les plus récents, plus particulièrement pour le mois le plus récent. Cela peut-être interprété par la non-validité de l'hypothèse linéarité (H2), où la transition du premier au deuxième mois n'est pas stable. D'où l'écart observé est justifié, puisque l'estimation du montant ultime de la sinistralité relative au mois le plus récent nécessite une estimation de transition du premier au deuxième mois.

Un écart de 320 145 TND sur un montant réel nécessaire de provision de 778 288 TND, soit 41,13%, peut provoquer des problèmes financiers au niveau de la compagnie d'assurance.

	IBNR par Mack CL	IBNR Réels	Erreur(%IBNR)
Pour 2021	458 143 TND	778 288 TND	41, 13%

TABLE 4.2 – Performance du modèle de Mack Chain Ladder

Les résultats présentés dans la Table 4.2 prouvent la nécessité d'une autre approche pour estimer correctement le montant des réserves que la compagnie d'assurance doit garder pour les IBNRs.

4.4 Modèle GLM

Pour cette approche, on va considérer le même triangle de données cumulées pour pouvoir comparer les résultats au modèle de Mack. A cet effet, l'application du modèle GLM nécessite de transformer le triangle en une table de données à trois colonnes qui sont :

- X_1 : Mois de survenance
- X_2 : Mois de développement
- Y : Montant payé

Pour la distribution de Y, on va choisir une Gaussienne parce qu'en comparant les résultats obtenus par les différents modèles, on remarque que l'AIC relatif à la famille gaussienne est plus petit. Et pour la fonction de lien, on va choisir l'identité qui est la fonction lien canonique de la loi normale.

Résultats du modèle GLM

Le modèle GLM a généré les montants ultimes des sinistralités relatives aux 12 mois de l'année 2021, soit les résultats présentés dans la Figure 4.7.

déja payé	Observé	GLM
620068.8376	620068.8376	620069.000000
661829.2450	661829.2450	672348.454545
754548.9144	755111.9524	758363.727273
555714.4083	555752.8923	599676.023569
549372.8436	551113.1376	587208.488847
609804.1090	611541.2730	643185.881704
442212.9070	445391.6300	497910.345990
530082.4290	532175.8000	574770.645990
678863.8730	682380.6630	707349.383490
630724.3810	652288.9550	631928.179786
587294.5080	720694.7470	659435.013119
91317.6880	701772.5030	590812.740392

FIGURE 4.7 – Résultats des estimation par le modèle GLM

Le total de la sinistralité relative à l'année 2021, estimé par le modèle GLM, est égal au montant suivant : 7 543 058 TND, soit un écart de +52 936 TND qui représente 0,707% du vrai montant ultime. Ainsi pour le montant de provision en retrouve le même écart.

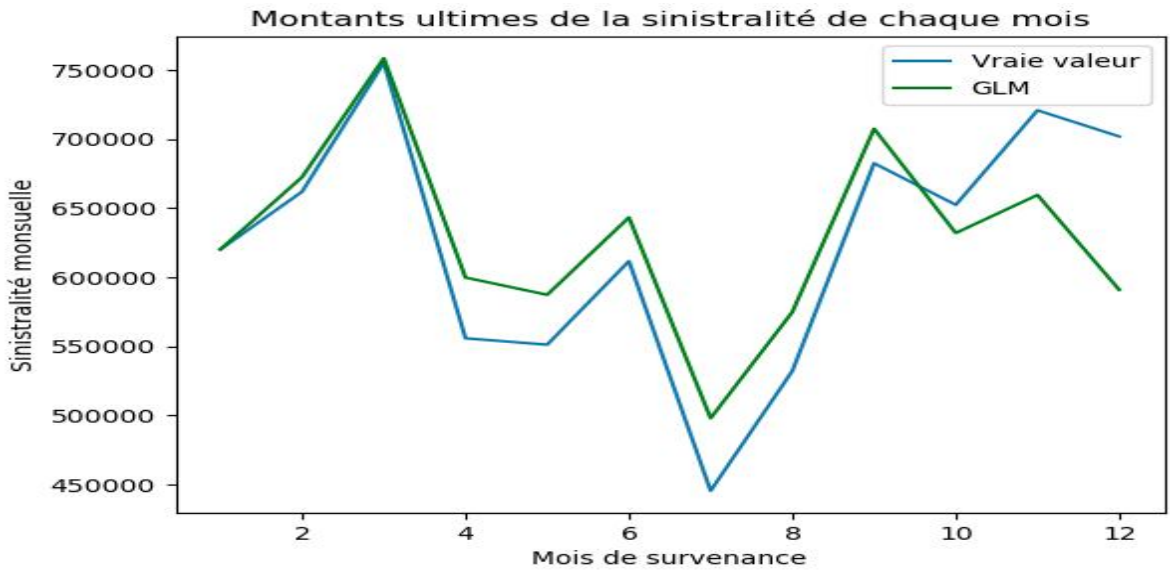


FIGURE 4.8 – Montants mensuels ultimes estimés par le modèle GLM

La Figure 4.8 montre que le modèle GLM surestime les montants des sinistralités des mois les plus anciens et sous-estime la sinistralité des mois le plus récents. En gros, les écarts ne sont pas très élevés même pour le mois le plus récent, qui est difficile à estimer par le modèle Chain Ladder de Mack.

	IBNR par le GLM	IBNR Réels	Erreur(%IBNR)
Pour 2021	831 224 TND	778 288 TND	6, 8%

TABLE 4.3 – Performance du modèle GLM

La Table 4.3 nous donne une indication sur taux d'erreur du modèle GLM par rapport à la vraie valeur.

Intervalle de confiance

Afin de construire un intervalle de confiance pour le montant de réserve à estimer, on a besoin d'une distribution des IBNRs. Pour cela, on va considérer les erreurs de déviation dû à la prédiction du triangle supérieur. On va ré-échantillonner, avec remise, 10 000 fois les erreurs et les redistribuer afin de générer des faux triangles.
Par la suite, on va appliquer 10 000 modèles sur les 10 000 triangles construits. Finalement, on va obtenir une distribution de 10 000 valeurs d'IBNR.

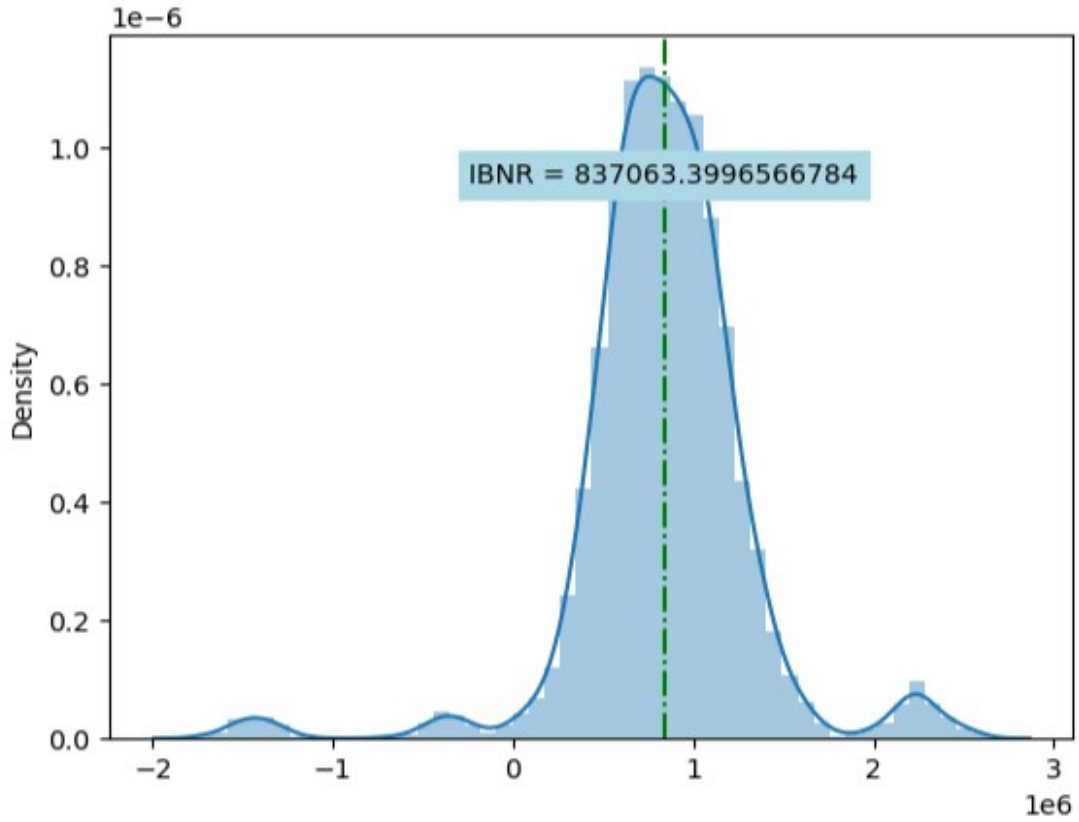


FIGURE 4.9 – Distribution des IBNRs estimés par le modèle GLM

La distribution des IBNRs présentée dans la Figure 4.9 semble être une distribution normale. L'intervalle de confiance pour les montants de provision est donc égal à :

$$IC_{IBNR}^{95\%} = 837\,063,39 \pm 9\,459,53.$$

4.5 Comparaison des performances

	Mack CL	GLM
$Erreur$(%IBNR)	$41,13\%$	$6,8\%$
$RMSE$	$81\,318,31$	$46\,574,38$

TABLE 4.4 – Comparaison des performances entre GLM et Mack Chain Ladder

Ici l'Erreur est définie par : $Erreur(\%) = \dfrac{|vraie\ valeur - valeur\ predite| \times 100}{vraie\ valeur}$. La Table 4.4 montre que le modèle GLM est bien plus performant que celui de Mack.

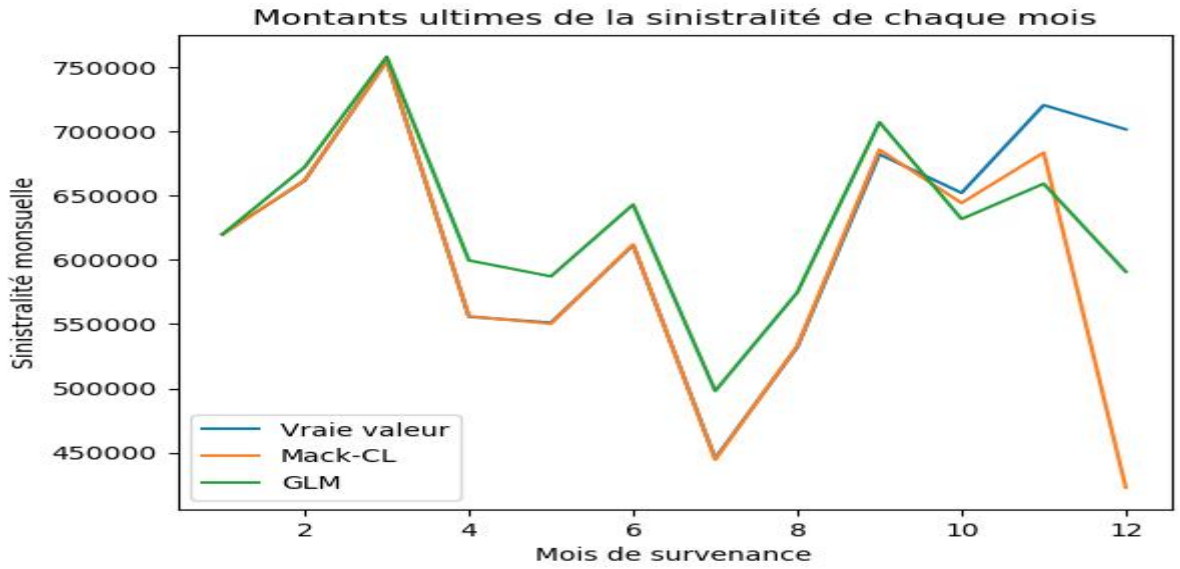


FIGURE 4.10 – Comparaison des estimations des IBNRs mensuels

On constate, dans la Figure 4.10, que le modèle Chain Ladder de Mack prédit efficacement les montants ultimes correspondants aux mois de survenance les plus anciens, alors que la prédiction pour le mois le plus récent est erronée. Le modèle GLM vient de réduire cette erreur et surestime légèrement les montants correspondants aux mois les plus anciens pour compenser le reste de l'erreur.

4.6 Conclusion

Notre but dans cette partie était de trouver une alternative pour l'estimation des provisions pour sinistres à payer dans le cas où les hypothèses de Mack pour l'application du modèle Chain Ladder ne sont pas valides. Finalement, l'approche GLM qu'on a proposé était plus robuste comparée au modèle de Mack. En effet, on l'a appliqué sur plusieurs autres données pour d'autres polices et elle a donné de bons résultats (Voir l'annexe Table 2).

Troisième partie

Détection de fraude

Chapitre 5

Apprentissage non-supervisé

5.1 Introduction

L'approche idéale pour des études de classification binaire, comme dans notre cas (Fraude ou non-fraude), est l'apprentissage supervisé qui consiste à apprendre auprès des données historiques et classifie les évènement à venir en se basant sur cet apprentissage [9]. Si les données nécessaires sont disponibles, un modèle de classification peut être appliqué directement pour faire la prédiction des évènements à classer.
Le problème dans le cas de détection des fraudes est que ces données ne sont pas toujours disponibles. Il faut noter que la rareté de ces évènements rend l'apprentissage des modèles de classification difficile.

L'apprentissage non-supervisé est une autre alternative lorsque les données labelisées ne sont pas disponibles. Cette approche ne suppose aucune information préalable sur les fraudeurs dans l'ensemble de données. Mais plutôt, elle détecte les comportements atypiques par rapport au reste. Il convient, bien sur, de noter qu'un tel comportement ne présente pas toujours une fraude. Le problème avec ce type de modèles est qu'ils sont considérés comme étant des boîtes noir (*Black Boxes*) où l'*output* ne soit pas explicables.

Le défi dans cette étude n'est pas seulement la détection de l'anomalie, mais aussi d'expliquer les raisons pour laquelle elle est considérée comme telle. Dans cette partie du projet, une forêt d'isolement (Isolation Forest) a été utilisée comme algorithme de base pour la détection des anomalies.

5.2 Modèles d'apprentissage non supervisé

Cette section va porter sur l'ensemble des modèles d'apprentissage non-supervisé qu'on va utiliser dans le processus de détection de la fraude. On va commencer par présenter la technique de réduction de la dimension ACP qui a pour but de visualiser les données dans un espace de 2D ou 3D. en suite nous présenterons les modèles de clustering (K-means et hiérarchique). Finalement, on terminera par expliquer le principe du modèle de détection d'anomalie Isolation Forest.

5.2.1 Analyse en Composantes Principales (ACP)

Lorsqu'on travaille avec des données de haute dimension, l'interprétabilité des données s'avère souvent difficile. Ainsi, le traçage des données en plus que 3D sera impossible. D'où l'importance de l'ACP qui est une technique de réduction de la dimensionnalité des données. Elle augmente l'interprétabilité tout en minimisant la perte de l'information. Elle permet de trouver les caractéristiques les plus significatives dans un ensemble de données et facilite le tracé des données en 2D et 3D.

Les composantes principales sont des combinaisons linéaires des variables initiales.

$$c^j = \sum_{i=0}^{n} a_i x^i \tag{5.1}$$

Les composantes principales sont non corrélées deux à deux. La représentation des individus sur le $j^{`eme}$ axe principal est donnée par

$$c^j = \begin{pmatrix} c_1^j \\ c_2^j \\ \vdots \\ c_n^j \end{pmatrix}. \tag{5.2}$$

Si on désire une représentation plane des individus, la meilleure sera celle réalisée grâce aux deux premières composantes principales, comme indiquée dans la Figure 5.1.

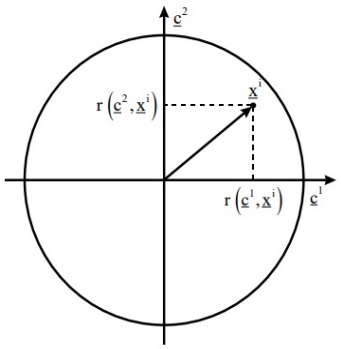


FIGURE 5.1 – Représentation des variables dans le nouvel espace [5]

Les **proximités** entre les variables initiales et les composantes principales sont mesurées par les covariances et surtout les corrélations. La corrélation entre la composante c^j et la variable x^i est notée $r(c^j, x^i)$.

5.2.2 K-means

L'algorithme des K-moyennes (ou K-means) est un algorithme de classification non-supervisé. Il regroupe en K groupes, l'ensemble d'individus. Ainsi, les individus se trouvant dans un même groupe ont des caractéristiques similaires.

Cet algorithme est basé sur la minimisation des distances entre les point. Il utilise essentiellement la distance euclidienne. Supposons que X et Y sont deux individus dans un un espace de n-dimension, $X = (x_1, x_2, .., x_n)^T$ et $Y = (y_1, y_2, .., y_n)^T$. La distance entre ces deux individus est égale à

$$D = ||Y - X|| \tag{5.3}$$

$$= \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5.4}$$

Pour arriver au résultat final :

— L'algorithme tire au hasard K centres $c_1^{(0)}$, $c_2^{(0)}$ ... $c_K^{(0)}$. On regroupe les individus autour de ces K centres, de sorte que la classe associée au centre $c_k^{(0)}$ est constituée des individus les plus proches de $c_K^{(0)}$ que tout autre centre ; on obtient ainsi la partition $C_1^{(0)}$, $C_2^{(0)}$ ... $C_K^{(0)}$.

— On calcule les centres de gravité $g_1^{(1)}$, $g_2^{(1)}$ ... $g_K^{(1)}$ des K classes, on effectue une deuxième partition autour des K centres. On obtient ainsi la partition $C_1^{(1)}$, $C_2^{(1)}$ ... $C_K^{(1)}$.

L'algorithme continu ainsi jusqu'à ce que la qualité de la partition mesurée par un critère convenablement choisi (i.e. l'inertie intra-classes) ne s'améliore plus.

5.2.3 Classification hiérarchique

La classification hiérarchique est l'une des méthodes non-supervisée la plus utilisée dans la classification des données multidimensionnelles. Cette méthode commence par un ensemble de points distincts, chaque point est considéré comme classe indépendante. Les deux classes les plus proches selon un critère donné sont regroupées. Cela se répète jusqu'à ce que tous les points appartiennent à une classe construite de manière hiérarchique. La structure hiérarchique finale est dite *dendrogramme*.

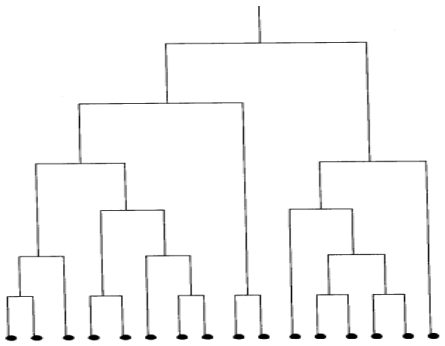


FIGURE 5.2 – Exemple de dendrogramme

Les critères de regroupement (ou d'agrégation) de deux classes souvent utilisés sont détaillés ci-dessous.

— Critère du lien minimum :

$$d(\mathcal{C}_1, \mathcal{C}_2) = \min_{\omega_i \in \mathcal{C}_1, \omega_{i'} \in \mathcal{C}_2} d(\omega_i, \omega_i') \tag{5.5}$$

— Critère du lien maximum :

$$d(\mathcal{C}_1, \mathcal{C}_2) = \max_{\omega_i \in \mathcal{C}_1, \omega_{i'} \in \mathcal{C}_2} d(\omega_i, \omega_{i'}) \tag{5.6}$$

— Critère de la moyenne :

$$d(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{n_1 n_2} \sum_{\omega_i \in \mathcal{C}_1} \sum_{\omega_{i'} \in \mathcal{C}_2} d(\omega_i, \omega_{i'}) \tag{5.7}$$

— Critère de *Ward*. L'inertie d'un nuage de points peut se décomposer comme suit :

$$\begin{aligned} I_{totale} &= I_{inter-classe} + I_{intra-classe} && (5.8) \\ &= \sum_k n_k \, d^2(g_k, g) + \sum_k \sum_{i \in \mathcal{C}_k} d^2(x_i, g_k) && (5.9) \end{aligned}$$

Lorsque l'on passe d'une partition en $k+1$ classes à une partition en k classes en regroupant deux classes en une seule l'inertie inter-classe diminue (l'inertie intra-classe augmente). Le critère de regroupement consiste à regrouper les deux classes pour lesquelles la perte d'inertie est la plus faible. Ceci revient à réunir les deux classes les plus proches en prenant comme distance entre deux classes la perte d'inertie inter-classe que l'on encourt en les regroupant :

$$d(\mathcal{C}_1, \mathcal{C}_2) = \frac{n_1 n_2}{n_1 + n_2} \, d^2(g_1, g_2) \tag{5.10}$$

où n_k et g_k sont, respectivement, le cardinal et le centre de gravité de la classe $\mathcal{C}_k$.

5.2.4 Isolation Forest : concept de base et principe de prédiction

Cet algorithme est l'algorithme non-supervisé le plus souvent utilisé pour la détection d'anomalie[11]. Le forêt d'isolement (Isolation Forest) est, comme le forêt aléatoire (*Random Forest*), basé sur plusieurs arbres de décision.

En phase d'apprentissage, chaque arbre de décision décompose l'espace de données en deux sous-arbres en fonction de valeurs arbitraires d'une variable choisie aléatoirement. Chaque sous-arbre suit le même processus jusqu'à atteindre une condition d'arrêt ;

(a) Si un nœud contient une seule observation.

(b) Si l'arbre atteint sa hauteur maximale.

Ce qui fait qu'une observation anormale sera facilement détectée, alors qu'une observation normale nécessitera plusieurs itérations pour être isolée (voir Figure 5.3).

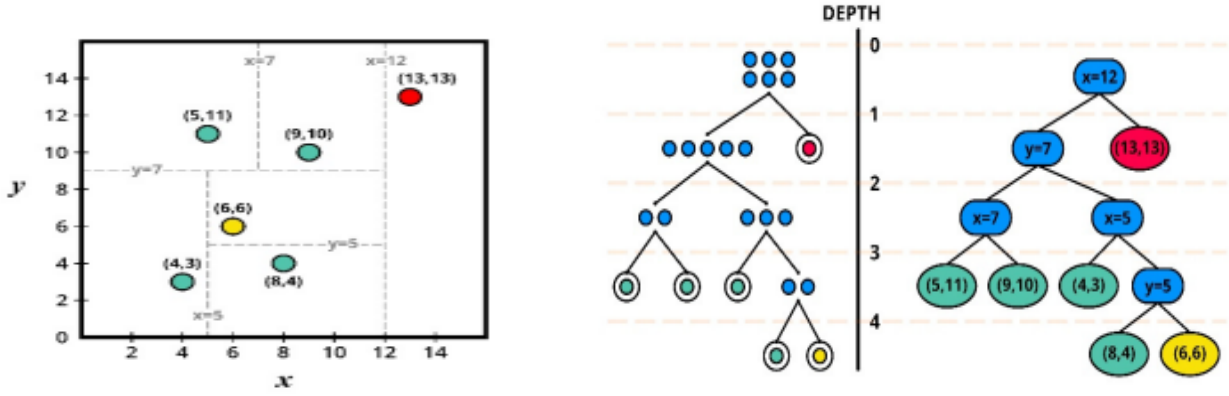


FIGURE 5.3 – Concept de base de l'Isolation Forest [8]

La Figure 5.3 illustre le concept de base de forêt d'isolement. Comme le montre la figure, un arbre de recherche binaire a été construit de manière aléatoire pour isoler chaque point de données. Dans la première figure, une seule division était suffisante pour séparer le point anormal $(13, 13)$ du reste des données, alors que quatre divisions étaient nécessaires pour séparer le point normal $(6, 6)$. Comme le suggère la figure, l'instance anormale est susceptible d'être plus proche du nœud racine que l'instance de données normales.

Les observations anormales sont susceptibles d'être plus proches du nœud principale. Donc la longueur du chemin suivi pour atteindre un noeud terminal sera considérée comme une mesure pour qualifier une observation comme normale ou anormale (voir Figure 5.3).

Dans la phase de prédiction, l'Isolaton Forest mesure le degré de l'anomalie en mesurant la longueur moyenne du chemin entre la racine et l'instance pour trouver la longueur de chemin attendue. Plus formellement, un score d'anomalie est donnée pour chaque observation pour la prise de décision. Le score d'atypisme d'une observation x est donné comme suit :

$$S(x, n) = 2^{-\frac{\mathbb{E}(h(x))}{c(n)}} \tag{5.11}$$

avec

- x : Une observation/un individu.
- $h(x)$: La longueur du chemin de x, c'est-à-dire le nombre total de séparations nécessaire pour soler l'individu x.
- $\mathbb{E}(h(x))$: La moyenne de $h(x)$ sur tous les arbres.
- n : La taille de l'échantillon dans les différents arbres de décision.
- $c(n)$: La longueur moyenne du chemin compte tenu de la taille de l'échantillon.

$$c(n) = 2 \times H(n-1) - 2 \times \frac{(n-1)}{n} \tag{5.12}$$

où $H(n-1)$ est estimée par la constante d'Euler 0.5772156649 :

$$H(n-1) = 1 + \frac{1}{2} + ... + \frac{1}{n-1} \tag{5.13}$$

$$= ln(n-1) + 0.5772156649 \tag{5.14}$$

Si le score d'anomalie est proche de 1, l'observation x est considérée comme anomalie, et si $S(x, n)$ est inférieur à 0.5, x est probablement une observation normale.

Après avoir estimé le score d'anomalie pour chaque individu de notre ensemble de données, nous pouvons les trier par ordre décroissant pour trouver les principales anomalies ayant les pires scores.

Comme on l'avait dit précédemment, ces équations ne sont pas suffisantes pour l'explication de la décision à propos de l'atypisme de l'observation. Ce qui nécessite un outil pour évaluer la contribution des différentes variables dans la décision.

5.3 Traitement des comportements aberrants

Dans cette section nous allons nous intéresser aux comportements aberrents détectés par l'algorithme Isolation Forest, défini précédemment, afin de mieux comprendre ces comportements pour pouvoir valider l'hypothèse de fraude sur ces observations. Dans ce contexte, nous allons exploiter des techniques de regroupement (*clustering*) avec l'apprentissage non-supervisé telles que la classification hiérarchique et l'algorithme des K-means et d'exploration de données avec la technique d'analyse en composantes principales(ACP). Ces algorithmes sont applicable dans notre cas puisque nous disposons des données quantitatives.

Les algorithmes de classification non-supervisés tels que l'*hiérarchique* et le K-means nécessitent de définir, en amont, le nombre de classes à définir. Pour cela, nous allons utiliser des critères tels que le critère de coude et la silhouette.

Critère de Silhouette

On définit la Silhouette d'un objet i par :

$$S(i) = \frac{b_i - a_i}{max(a_i, b_i)} \tag{5.15}$$

où :

- a_i est la distance moyenne entre X_i et tous les autres objets de la classe $C(i)$ (La classe à laquelle appartient l'objet X_i) :

$$a_i = \frac{1}{|C(i)|} \sum_{X \in C(i)} d(X_i, X) \tag{5.16}$$

- b_i est la distance moyenne entre X_i et les objets de la classe la plus proche de X_i :

$$b_i = \min_{C \neq C(i)} \frac{1}{|C|} \sum_{X \in C} d(X_i, X) \tag{5.17}$$

Notons que $d(X, Y)$ est la distance euclidienne entre les deux objets X et Y.

- L'**indice de Silhouette** de validation d'une partition (i.e. du bon nombre de classes d'une partition) est la moyenne des Silhouettes de tous les objets.
- La meilleur partition est celle dont le nombre de classes correspond à :

$$arg\max_{k} S(i), \quad k = 2, ..., K \tag{5.18}$$

Critère de Coude

L'une des méthodes usuelles pour choisir le nombre de classes est de lancer K-means avec différentes valeurs de K et de calculer la variance des différents *clusters*. La variance est la somme des distances entre chaque *centroid* d'un *cluster* et les différentes observations du même *cluster*. Ainsi, on cherche K de manière que les classes retenues minimisent la distance entre leurs centres (*centroids*) et les observations dans la même classe. On parle donc du minimisation des distances **intra-classe**.

La variance des clusters se calcule ainsi :

$$\mathbb{V} = \sum_{k=1}^{K} \sum_{x_i \in C_k} d^2(c_k, x_i) \tag{5.19}$$

où c_i est le centre du *cluster* (appelé *centroid*).

Généralement, en représentant graphiquement les différents nombres de clusters K en fonction de la variance, on observe un schéma graphique similaire à celui illustré ci-dessous : Généralement, le

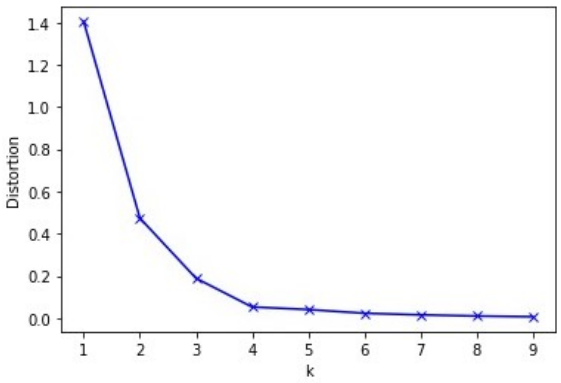


FIGURE 5.4 – Critère de coude

point du coude correspond au nombre de clusters à partir duquel la réduction de la variance devient moins significative. C'est le point ou K est le nombre optimal de *clusters*. Dans ce cas, le coude peut être représenté par K ayant une valeur de 2 ou 3.

5.4 SHAP pour l'interprétation des modèles de ML (*XAI*)

L'explication des résultats (*output*) des modèles d'apprentissage automatique peut permettre l'amélioration de l'interprétabilité de ces résultats, ce qui aide à la prise de décision.

Expliquer la prédiction d'un modèle peut être une tâche difficile. Ceci est particulièrement vrai pour les modèles qui utilisent des méthodes d'ensemble (i.e. les modèles qui combinent les résultats de nombreux modèles d'apprentissage de base indépendants).

Cette partie de notre projet vise à expliquer la prédiction, par l'algorithme Isolation Forest, de chaque observation en mesurant la contribution de chaque variable à cette prédiction.
Supposons que nous souhaitons expliquer la prédiction d'un *input* x par un modèle $f(x)$. Si nous pouvons représenter l'unique prédiction $f(x)$ par la somme de l'importance individuelle des caractéristiques ϕ_i, nous pouvons alors expliquer la prédiction $f(x)$ comme suit :

$$f(x) = g(x') \tag{5.20}$$

$$= \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{5.21}$$

où :

- x' est un *input* simplifié qui correspond à l'*input* de base x par le biais d'une fonction de mise en correspondance : $x = h_x(x')$. Ici x' est un vecteur binaire $x' \in \{0,1\}^M$ où 1 signifie que la i^{eme} composante est présente et 0 qu'elle est absente.
- ϕ_0 est la valeur de base lorsque toutes les caractéristiques d'entrée sont manquantes : $\phi_0 = g(0)$.
- M est le nombre d'*inputs* simplifiés
- f : modèle à expliquer.
- g : modèle explicatif

Ce genre d'explication est dit *additive feature attribution method*. Plusieurs méthodes ont été proposées pour trouver une bonne fonction d'approximation g.

SHAP ou (*SHapley Additive exPlanations*) est une approche, proposée par Lundberg et Lee en 2016, qui vise à évaluer l'importance de chaque caractéristique dans une prédiction spécifique. Cette approche utilise les valeurs de Shapley, qui sont basées sur le concept de la théorie des jeux, établi par Lloyd Shapley en 1953. Le problème est décrit comme suit : « **Un groupe de personnes ayant des compétences différentes participe à une compétition pour obtenir des récompenses. Comment répartir équitablement ces récompenses entre tous les joueurs ?** »

Soit N joueurs. Pour une coalition de joueur S, $f(S)$ est la somme des récompenses que les membres de S peuvent obtenir en coopérant. Selon Shapley, le gain qu'un joueur i obtient dans le cadre d'un jeu de coalition (f, N) :

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!}(f(S \cup \{i\}) - f(S)) \tag{5.22}$$

La valeur de Shapley est la contribution marginale moyenne d'un joueur dans toutes les coalitions possibles, où $(f(S \cup \{i\}) - f(S))$ est la contribution marginale du joueur i compte tenu de la coalition actuelle S. La valeur de Shapley peut-être interprétée comme étant la moyenne de cette contribution marginale sur toutes les permutations possibles.

$$\phi_i = \sum_{\text{Coalitions sans } i} \frac{\text{Contribution marginale de i dans } S}{\text{Nombre de coalitions sans } i \text{ de cette taille}}. \tag{5.23}$$

La valeur de Shapley ϕ_i doit satisfaire les quatre propriétés suivantes :

1. **Efficience** : On veut distribuer toutes les récompenses, donc la somme de toutes les valeurs de Shapley doit être égale au montant total $f(N) = \sum_i \phi_i$.
2. **Symétrie** : Si deux joueur ont la méme contribution, alors leurs récompenses devraient être les mêmes.

   $$f(S \cup \{i\}) = f(S \cup \{k\}) \qquad alors \qquad \phi_i = \phi_k.$$

3. **Dummy** : La valeur de Shapley d'un joueur qui ne contribue pas à la récompense doit être nulle. Dans notre cas, un tel joueur est une caractéristique qui ne modifie pas la prédiction.

   $$f(S \cup \{i\}) = f(S) \qquad alors \qquad \phi_i = 0.$$

4. **Linéarité** : Si le même groupe joue deux parties, les récompenses de chaque joueur pour les deux parties doivent être égales à la somme des récompenses de la première et de la deuxième partie.

   $$\phi(f + F) = \phi(f) + \phi(F).$$

Les valeurs de Shapley peuvent être utilisées dans le contexte de l'apprentissage automatique et de l'interprétabilité en supposant que chaque valeur de caractéristique de l'instance est un joueur dans un jeu où la prédiction est le gain. Par conséquent, la structure d'attribution des caractéristiques additives (*Additive Feature Attribution*) pour les valeurs de Shapley est la suivante :

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{5.24}$$

avec

$$\phi_i(f,x) = \sum_{z' \subseteq x' \setminus \{i\}} \frac{|z'|!\,(M-|z'|-1)!}{M!}(f(z' \cup i) - f(z')) \tag{5.25}$$

où z' est un vecteur binaire représentant le sous-groupe S des variables incluses dans le modèle.

Naturellement, SHAP satisfait les quatre propriétés mentionnées précédemment puisqu'elle est basée sur les valeurs de Shapley. Cependant, Lundberg a mis en évidence trois autres propriétés que le SHAP devrait satisfaire. Ces propriétés sont :

5. **Précision locale** : Le modèle à expliquer f est approximé par l'*input* x, alors le modèle d'explication devrait au moins correspondre à la sortie du modèle f avec l'*input* simplifiée x'.

$$f(x) = g(x') \tag{5.26}$$
$$= \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{5.27}$$

   Le modèle d'explication $g(x')$ correspond au modèle original $f(x)$ lorsque $x = h_x(x')$.

6. **Absence** : Les caractéristiques manquantes ne devraient pas avoir d'impact.

$$x'_i = 0 \qquad \Longrightarrow \qquad \phi_i = 0$$

7. **Cohérence** : Si la contribution d'un *input* simplifié augmente ou reste la même en raison du changement de modèle, l'attribution de l'*input* doit être cohérente avec le changement. Supposons que $f_x(z') = f(h_x(z'))$ et que $z' \setminus i$ veut dire que $z'_i = 0$, pour deux modèles f et f' satisfaisant :

$$f_x(z') - f_x(z' \setminus i) \geq f'_x(z') - f'_x(z' \setminus i)$$

   pour tout $z' \in \{0,1\}^M$, alors :

$$\phi_i(f,x) \geq \phi_i(f',x)$$

D'après [12], un seul modèle, g, d'explication possible qui satisfait l'*additive feature attribution method* et les propriétés (5–7) :

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M-|z'|-1)!}{M!}(f_x(z') - f_x(z' \setminus i)) \tag{5.28}$$

avec $|z'|$ est le nombre d'éléments non nuls de z', et $z' \subseteq x'$ sont des vecteurs où les éléments non-nuls sont un sous ensemble des éléments non-nuls de x'.

En agrégeant les valeurs de Shapley de chaque instance, l'interprétation du modèle global sera aussi possible.

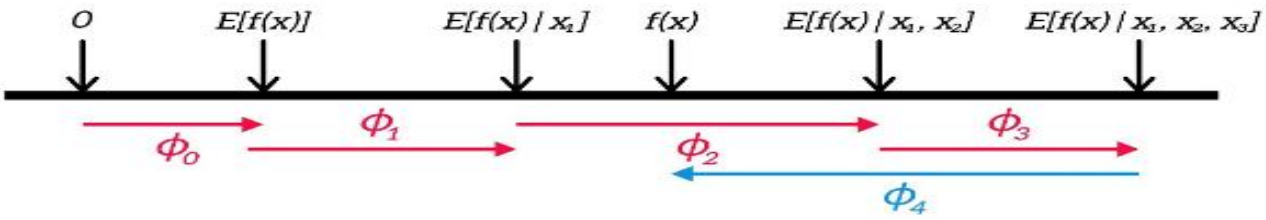


FIGURE 5.5 – Explication avec l'approche SHAP d'un modèle à quatre variables [11]

La Figure 5.5 illustre comment l'approche SHAP explique la prédiction d'un modèle de ML à quatre variables. Elle explique le chemin suivi pour passer de la valeur de base, $\phi_0 = \mathbb{E}[f(x)]$, et arriver au résultat du modèle (*output*).

5.5 Conclusion

Cette section était une opportunité pour comprendre les principes de la détection d'anomalies via l'algorithme Isolation Forest, ainsi que les autres modèles classiques d'apprentissage non-supervisé.

La compréhension des principes fondamentaux de l'approche SHAP, qui vise à expliquer les résultats des modèles de Machine Learning, nous permet l'ouverture sur des opportunités d'utilisation des modèles, précédemment dits «Black Boxes», dans des domaines sensibles, tel que l'assurance (qui exige de la transparence). D'où la nécessité de l'utilisation des modèles interprétables «Cristal Boxes».

En pratique, Nous avons utilisé l'algorithme Isolation Forest pour isoler les aberrations. Ensuite, nous les avons expliqué par l'approche SHAP. Finalement, nous les avons classé à l'aide de l'algorithme K-means qui montre des performances meilleurs que la classification hiérarchique. Ceci est présenté dans le sixième chapitre.

Chapitre 6

Exploration des résultats

Dans ce chapitre, nous allons explorer et discuter la fiabilité des résultats générés par nos modèles d'apprentissage non-supervisé.

La détection des comportements atypiques se fait par rapport à toute la population où les comportements doivent être relativement homogènes. Le comportement atypique sera constaté indépendamment de la catégorie de la population à laquelle il appartient.

Pour ce chapitre, on va exposer les resultats obtenus par l'application des modèles sur une population formée que de deux polices (groupes assurés) pour des raisons de confidentialité.

6.1 Isolation Forest

Pour l'implémentation de l'algorithme Isolation Forest, nous devons spécifier les variables explicatives sur lesquelles le modèle mesure le niveau d'atypisme d'un comportement. Ainsi, nous devront spécifier le taux de contamination.

En se basant sur des études faites précédemment, le taux de fraude en assurance est autour de 10% par rapport à la sinistralité totale. Nous allons prendre considérer un taux de contamination de 15% pour de modèle parce que ses résultats ne seront pas tous acceptés [1]. Le modèle retourne une label, pour chaque observation : "-1" pour "Anomalie" et "1" pour "Normale".

1. Sources mentionnées dans l'introduction de ce mémoire

6.1.1 Principales contributions

pour des dimensions supérieurs à trois, il est impossible de visualiser, et même d'imaginer, les données dans l'espace. D'où la nécessite de la réduction des dimensions. Pour cela, nous allons utiliser l'ACP qui est un outil de réduction de dimensions et d'exploration des données.

Résultats de l'ACP

Après avoir effectuer une ACP sur l'ensemble de données, on trouve les axes principaux qui expliquent la variabilité des données comme suit :

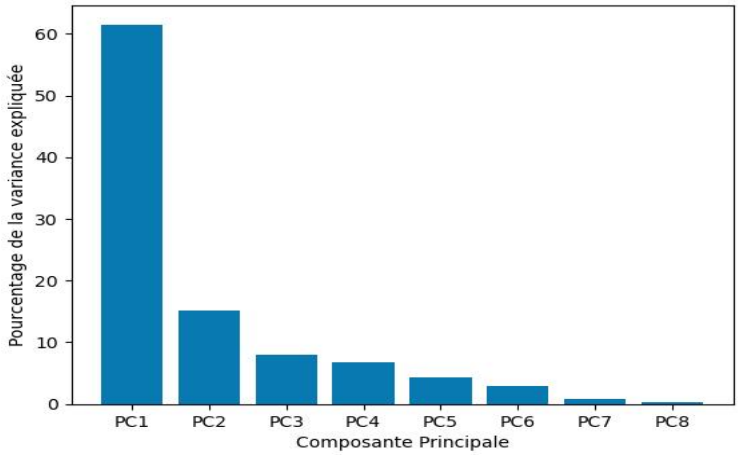


FIGURE 6.1 – Pourcentage de la variance expliquée par les composantes principales

La Figure 6.1 montre que les trois premières composantes principales expliquent plus que 80% de la variabilité des données.

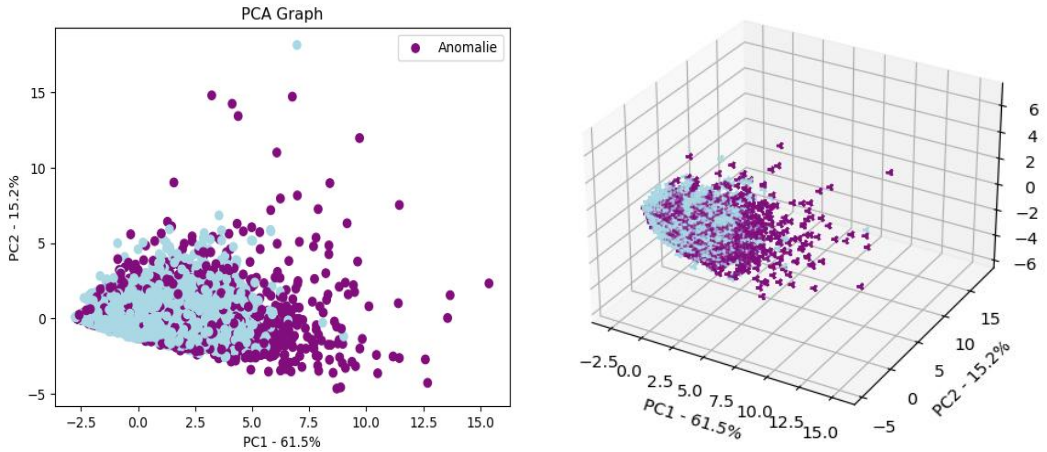


FIGURE 6.2 – Distribution des données par rapport aux axes principaux

On constate dans la Figure 6.2 que les anomalies détectées par l'algorithme sont distribuées avec dispersion par rapport aux axes principaux, d'où la nécessité d'expliquer pour chaque individu les raisons de labellisation (Anomalie ou Normale). Pour cela, nous utilisons l'approche SHAP.

6.1.2 SHAP

Pour justifier les résultats générés par le modèle aux preneurs de décision, nous allons examiner les contributions des différentes variables dans chaque prédiction. Considérons deux observations : comportement atypique et comportement normal.

Comportement atypique

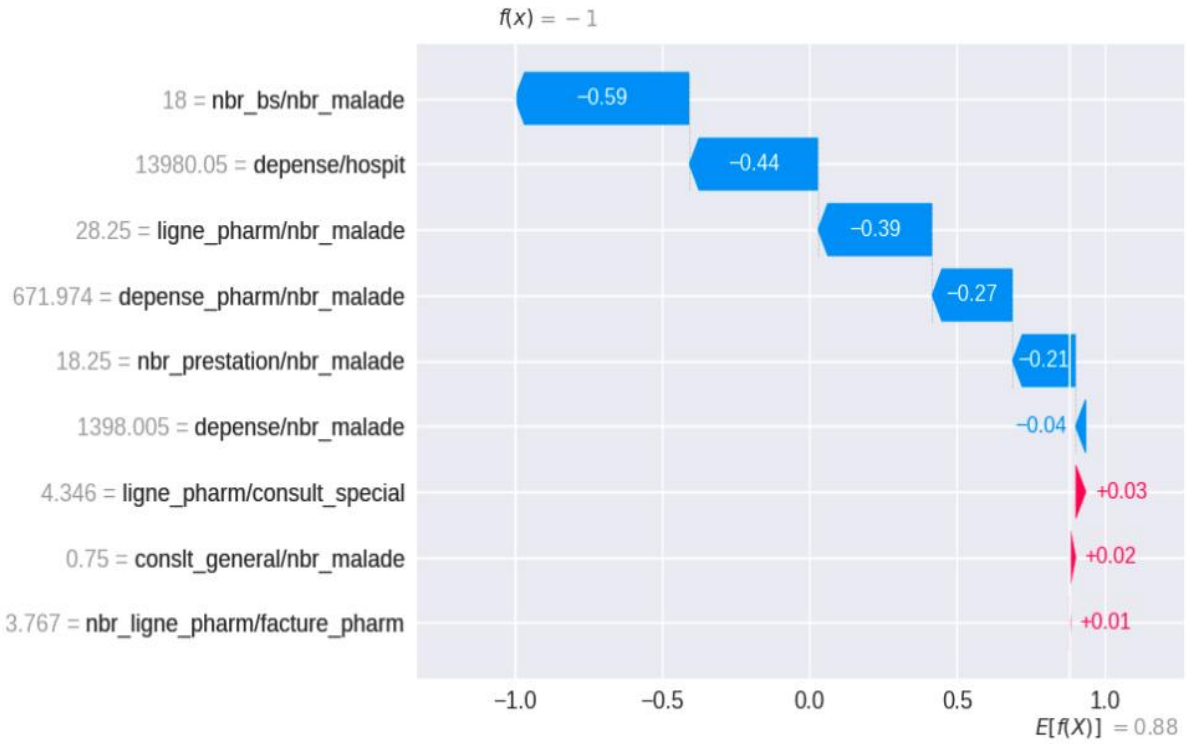


FIGURE 6.3 – Valeurs de SHAP pour une anomalie

La Figure 6.3 explique la contribution des différentes variables à la classification de ce comportement comme étant atypique en passant de la valeur de base $\mathbb{E}(f(X)) = 0,88$ à l'*output* $f(X) = -1$. Cette figure montre que `Nombre de bs par rapport au nombre de malade dans la famille`, `Dépense par rapport au nombre d'hospitalisation`, `Nombre de lignes pharmacie par rapport au nombre de malade dans la famille`, sont ceux qui contribuent le plus à cette classification. Pour les va-

riables `Nombre de lignes pharmacie par rapport au nombre de consultations généraliste`, `Consultation spécialistes par rapport au nombre de malades dans la famille` correspondent à des comportements normaux, sauf qu'elles n'étaient pas suffisantes pour classer cet individu comme normal.

Comportement normal

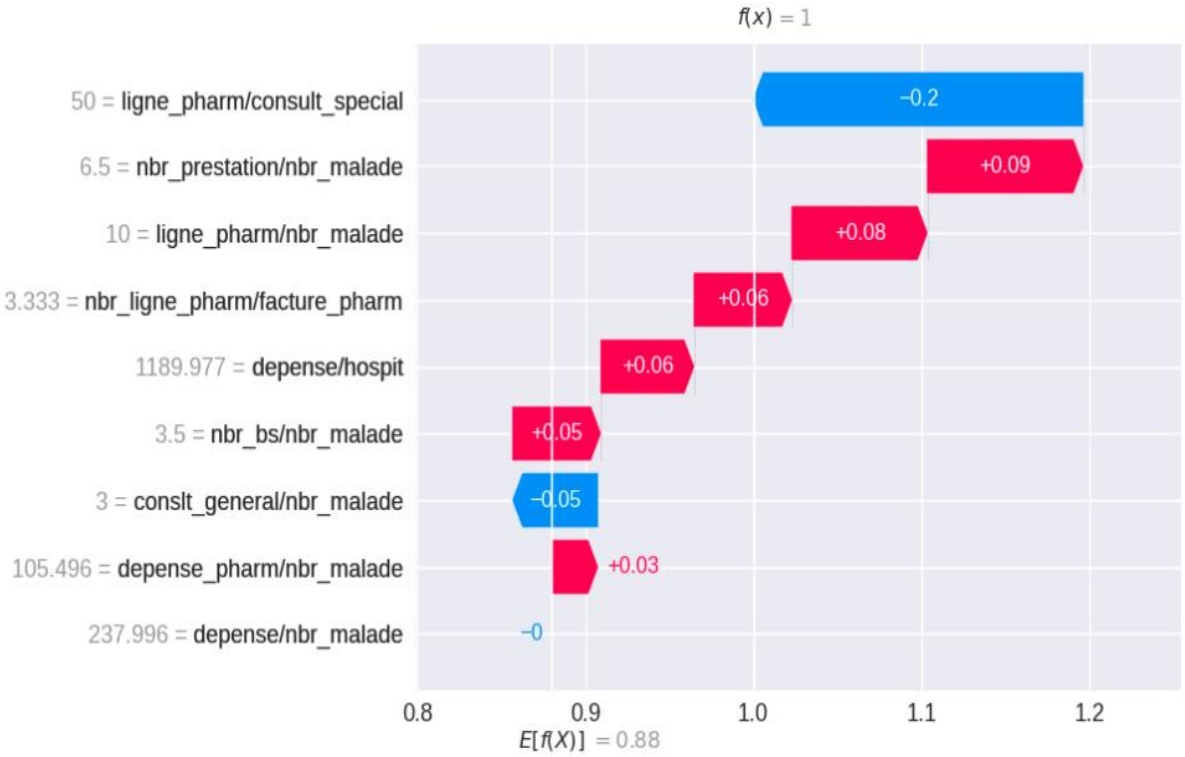


FIGURE 6.4 – Valeurs de SHAP pour un comportement normal

La Figure 6.4 nous montre que les variables `Nombre de prestations par rapport au nombre de malade dans la famille`, `Nombre de lignes pharmacie par rapport au nombre de malades dans la famille`, `Dépense pharmacie par rapport au nombre de malades dans la famille`, dont les valeurs sont inférieures à celles du comportement atypique, contribuent à la classification de cet individu comme normal. La valeur de la variable `Nombre de lignes pharmacie par rapport au nombre de consultations spécialistes` est relativement élevée, mais cette variable n'était pas suffisante pour classer cet individu comme anomalie.

Interprétation globale

En étudiant la distribution des valeurs SHAP pour toutes les observation, il est possible de définir la tendance générale de chaque variable dans la classification des individus.
La figure suivante montre l'interprétation globale des valeurs SHAP.

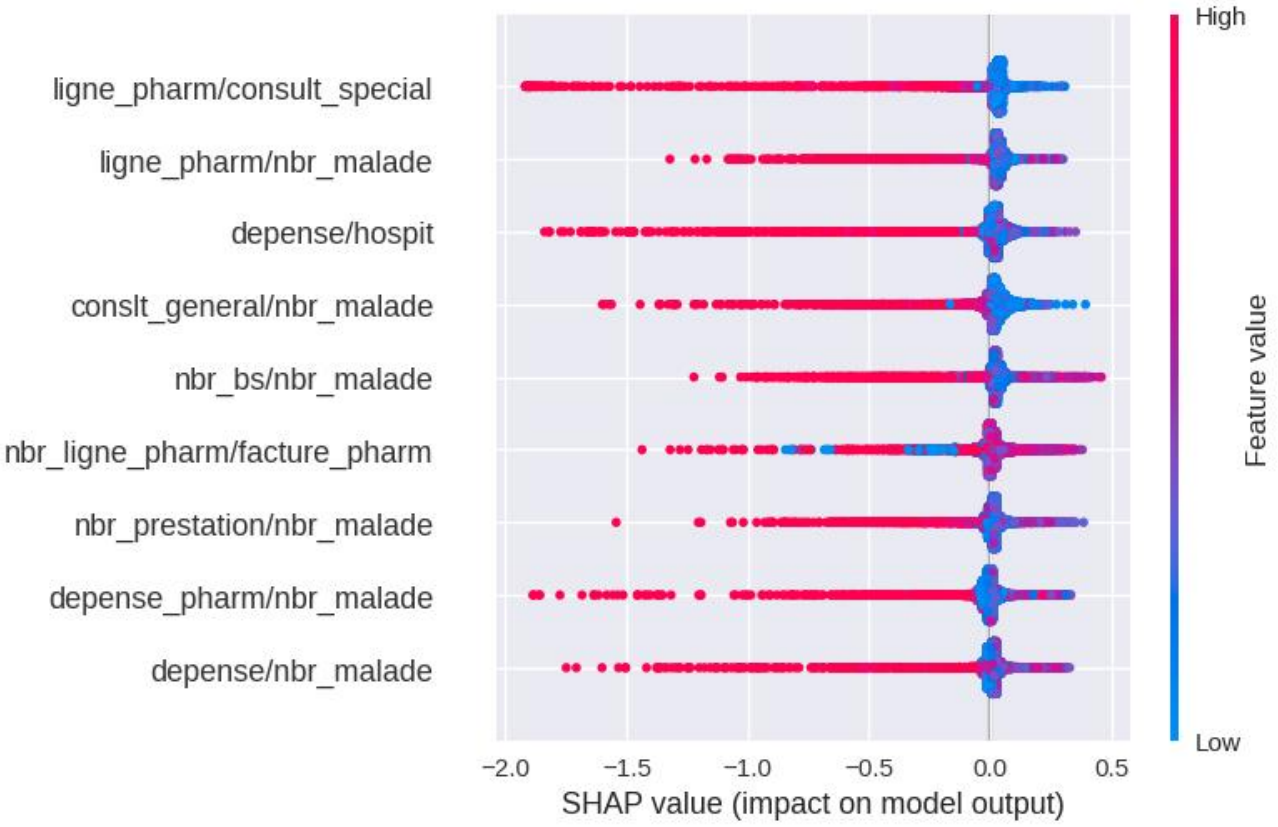


FIGURE 6.5 – Interprétation globale des valeurs SHAP

Le point dans chaque variable correspond à la valeur SHAP pour une observation donnée. Cependant, on doit se concentrer sur la partie gauche, car on veut expliquer les observations atypiques extrêmes.
La Figure 6.5 indique que, pour la majorité des variables utilisées, les grandes valeurs contribuent, dans la plus part des cas, à classer l'observation en tant qu'anomalie. Cela est dû au fait que la plus part des valeurs des différentes variables sont faibles. Et pourtant, on se retrouve avec des cas où les valeurs faibles sont considérées comme aberrations. Cela est dû à ce qu'on l'appelle *interactions entre les joueurs* dans le principe de la théorie des jeux.

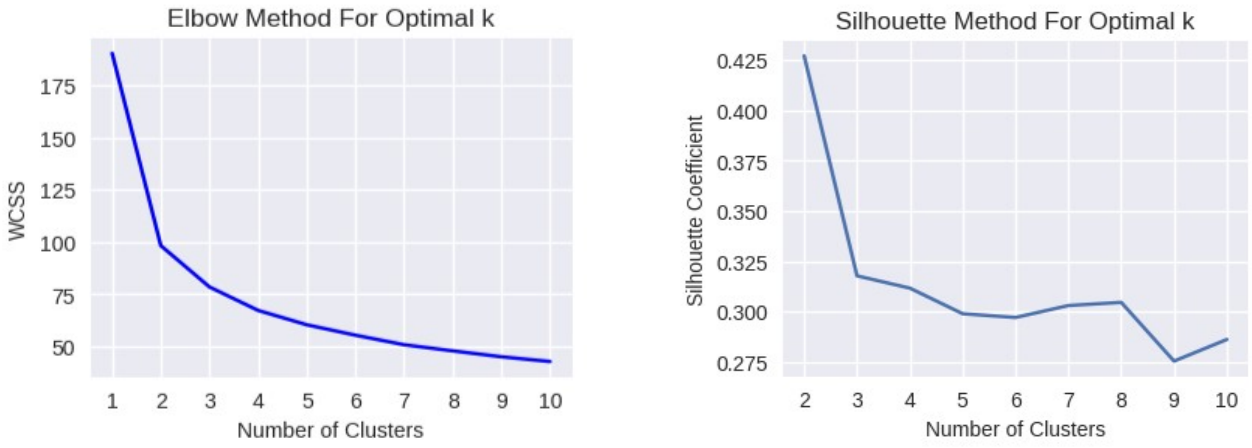
Bien que cette analyse nous a permis d'expliquer l'anomalie d'un point de vue local et global, l'utilisateur peut vouloir expliquer l'anomalie en trouvant des comportements communs à partir des anomalies, c'est-à-dire en regroupant les anomalies en fonction de leur similitude.

6.2 Partitionnement des anomalies

Afin de mieux comprendre les comportements atypiques, nous avons essayé de définir des groupes de comportements similaires à partir des aberrations. Au moins deux approches classiques d'apprentissage non-supervisé sont souvent utilisées pour classifier les données non labellisées qui sont, la classification hiérarchique et K-means. Dans cette section, nous allons présenter les résultats de l'algorithme K-means qui a donné de meilleures performances. Le critère de Silhouette pour K-means est plus grand que pour la classification hiérarchique, soit 0,41 pour le K-means et 0,32 pour la classification hiérarchique. Les résultats de la classification hiérarchique sont présents dans l'annexe.

Choix du nombre de groupes

En premier temps, nous allons choisir le nombre de *clusters*.



(a) Critère de coude

(b) Critère de Silhouette

FIGURE 6.6 – Critères du choix du nombre de groupes

Selon le critère de coude (Figure 6.6a), le coude correspond à deux classes, ainsi pour le critère de Silhouette (Figure 6.6b), la valeur maximale de Silhouette correspond à deux classes. Ainsi le nombre de classe optimal choisi est deux.

Exploration des résultats

Pour pouvoir visualiser nos résultats dans un espace de dimensions représentables, on a effectué une ACP sur les aberrations détectées par l'Isolation Forest. Ceci est fait dans le but de réduire les dimensions.

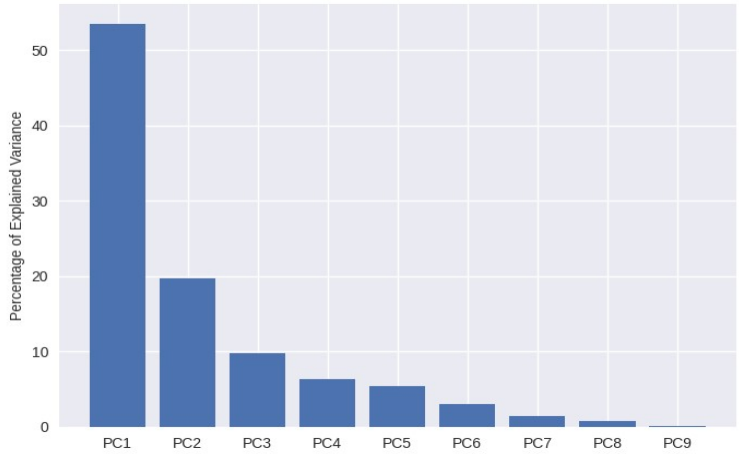


FIGURE 6.7 – Pourcentage de la variance expliquée par les composantes principales (anomalies)

La Figure 6.7 montre que les trois premiers axes principaux expliquent plus que 80% de la variabilité des données. La Table 6.1 montre les premières composantes principales (CP) et les variables qui les expliquent le plus.

Variable	CP1	Variable	CP2
`nombre de prestations`	0,4415	`hospitalisation`	0,7315
`nombre de bulletins de soin`	0,427	`dépense`	0,479
`nombre de factures pharmacie`	0,421	`nombre de bs labo`	0,392
`dépense`	0,353	`nbr de consultation généralistes`	-0,380

TABLE 6.1 – Corrélations entre les variables et les composantes principales

La Table 6.1 nous renseigne sur les corrélations entre les deux premiers axes principaux et les variables les plus corrélées avec chaque axes. Ceci nous permettra d'interpréter les groupes d'individus résultant de la classification par la méthode K-means. Les résultats de l'ACP montrent que :

- ✎ Le premier axe principal est corrélé avec les variables `nombre de prestations`, `nombre de bulletins de soin`, `nombre de factures pharmacie` et `dépense`.

✎ Le deuxième axe principal est corrélé avec les variables `hospitalisation`, `dépense`, `nombre de bulletins de soin labo` et `nombre de consultations généralistes`.

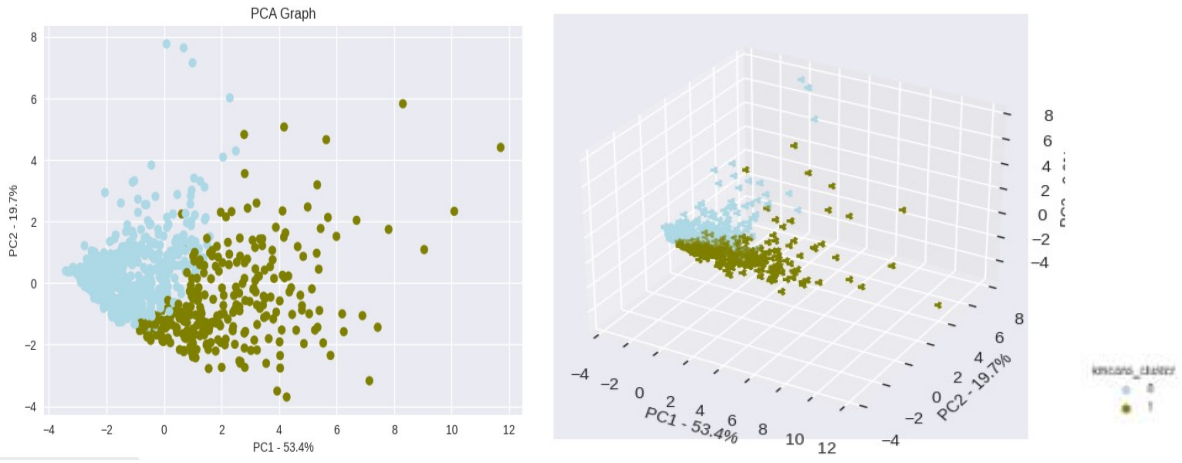


FIGURE 6.8 – Distribution des aberrations par rapport aux axes principaux

La Figure 6.8 et la Table 6.1 nous permettent d'interpréter les deux groupes obtenus par K-means. Un premier groupe d'individus (classe 1 en vert) qui tendent à consommer beaucoup en terme de nombre de prestations, bulletins de soin et factures pharmacie. On voit aussi que la majorité de ces derniers ont un taux d'hospitalisation faible. Ce sont les profils à retenir, après avoir consulté les experts métier.

Un deuxième groupe d'individus (classe 0 en bleu) qui ne dépensent pas beaucoup en matière de fréquence des sinistres en présence de quelque profils qui ont un taux d'hospitalisation élevé, ce qui reflète l'augmentation de leurs dépenses.

6.3 Conclusion

Dans ce chapitre, nous avons présenté les résultats de l'apprentissage non-supervisé, qu'on a appliqué afin de distinguer les profils atypiques parmi tous les assurés. L'algorithme Isolation Forest nous a aidé à isoler les profils aberrants. L'explication de cette classification était faite à l'aide de l'approche SHAP. Cette approche nous a indiqué que la plus part des cas, une variable à grande valeur contribue, généralement, à classer l'individu comme anomalie. On se retrouve avec quelques exceptions où une variable à valeur faible contribue à classer un individu comme anomalie ou le contraire. Cela est dû à l'effet de l'interaction entre les différentes variables. Finalement, nous avons terminer par regrouper les anomalies en deux classes.

Conclusions et Perspectives

Le domaine de l'assurance maladie est confronté à des défis importants qui mettent à l'épreuve sa capacité à garantir un accès équitable aux soins de santé, à fournir des prestations de qualité et à assurer la viabilité financière du système. L'un des plus grands challenge pour les compagnies d'assurance est d'estimer efficacement le risque pour pouvoir l'indemniser. Un deuxième enjeu majeur pour les assureurs est de faire face à la fraude.

Ce présent projet, élaboré dans le cadre de gestion de l'assurance maladie, a deux objectifs : estimer les provisions pour sinistres à payer et isoler les profils aberrants.

Dans un premier temps, nous avons présenté une approche innovante qui a pour objectif d'améliorer la qualité d'estimation des provisions pour sinistres à payer (PSAP). Pour cela, nous avons rempli le triangle de liquidation en se basant sur les modèles linéaires généralisés (GLM) en plus de procéder avec la méthode classique de Chain Ladder.
L'idée était de transformer le triangle de liquidation en trois colonnes : deux variables explicatives contenant les dates de survenance et les dates de paiement des sinistres, et une variable cible contenant les montants payés. Un test sur un an à été fait pour valider l'utilité de l'approche.

Il était important, avant de se lancer dans l'application du modèle GLM, de tester la méthode de Mack Chain Ladder. Cette méthode à démontré des faibles performances vu l'invalidité de ses hypothèses. Finalement, l'alternative que nous avons proposé était satisfaisante et plus robuste que la méthode de Mack Chain Ladder.

Dans un deuxième temps, nous avons exploité l'algorithme d'apprentissage non-supervisé *Isolation Forest* afin d'identifier les consommations aberrantes en matière de sinistres. Cette approche nous a permis de détecter les comportements atypiques dans un espace multidimensionnel. Afin d'assurer

une interprétation des anomalies, i.e. la raison de la classification d'une observation comme atypique, nous avons combiné l'*Explainable AI (XAI)* avec l'algorithme *Isolation Forest*. Dans notre étude, nous avons exploité l'approche SHAP, qui est mathématiquement robuste, afin de mesurer les contributions marginales des différentes variables à chaque prédiction. Finalement, nous avons terminé par définir deux groupes qui caractérisent les tendances des comportements atypiques : un premier groupe d'individus qui ont une tendance de consommation anormale et un second groupe contenant les individus à faible consommation.

Nos futurs projets viseront à effectuer des analyses longitudinales des comportements des assurés. Cela consiste à suivre la conduite des assurés tout au long d'une période considérée. Nous viserons aussi à exploiter l'*Explainable AI (XAI)* et la méthode SHAP pour expliquer des algorithmes tels que XGBoost et LSTM, qui sont des *Black Boxes* utiles pour le calcul des provisions.

Annexes

Generalized Linear Model Regression Results

Dep. Variable:	value	No. Observations:	78
Model:	GLM	Df Residuals:	55
Model Family:	Gaussian	Df Model:	22
Link Function:	identity	Scale:	1.9864e+09
Method:	IRLS	Log-Likelihood:	-932.03
Date:	Mon, 29 May 2023	Deviance:	1.0925e+11
Time:	12:28:34	Pearson chi2:	1.09e+11
No. Iterations:	3	Pseudo R-squ. (CS):	1.000
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.206e+05	1.96e+04	6.165	0.000	8.22e+04	1.59e+05
origin[T.1]	5.228e+04	1.9e+04	2.751	0.006	1.5e+04	8.95e+04
origin[T.10]	3.937e+04	3.59e+04	1.098	0.272	-3.09e+04	1.1e+05
origin[T.11]	-2.926e+04	4.87e+04	-0.601	0.548	-1.25e+05	6.61e+04
origin[T.2]	1.383e+05	1.97e+04	7.019	0.000	9.97e+04	1.77e+05
origin[T.3]	-2.039e+04	2.04e+04	-0.998	0.318	-6.04e+04	1.97e+04
origin[T.4]	-3.286e+04	2.13e+04	-1.545	0.122	-7.45e+04	8816.447
origin[T.5]	2.312e+04	2.22e+04	1.039	0.299	-2.05e+04	6.67e+04
origin[T.6]	-1.222e+05	2.35e+04	-5.209	0.000	-1.68e+05	-7.62e+04
origin[T.7]	-4.53e+04	2.5e+04	-1.811	0.070	-9.43e+04	3723.446
origin[T.8]	8.728e+04	2.72e+04	3.214	0.001	3.41e+04	1.41e+05
origin[T.9]	1.186e+04	3.04e+04	0.391	0.696	-4.76e+04	7.13e+04
development[T.10]	4.945e+05	3.04e+04	16.291	0.000	4.35e+05	5.54e+05
development[T.11]	4.942e+05	3.59e+04	13.783	0.000	4.24e+05	5.64e+05
development[T.12]	4.995e+05	4.87e+04	10.263	0.000	4.04e+05	5.95e+05
development[T.2]	3.944e+05	1.9e+04	20.751	0.000	3.57e+05	4.32e+05
development[T.3]	4.637e+05	1.97e+04	23.534	0.000	4.25e+05	5.02e+05
development[T.4]	4.664e+05	2.04e+04	22.822	0.000	4.26e+05	5.06e+05
development[T.5]	4.685e+05	2.13e+04	22.031	0.000	4.27e+05	5.1e+05
development[T.6]	4.711e+05	2.22e+04	21.181	0.000	4.28e+05	5.15e+05
development[T.7]	4.763e+05	2.35e+04	20.310	0.000	4.3e+05	5.22e+05
development[T.8]	4.797e+05	2.5e+04	19.177	0.000	4.31e+05	5.29e+05
development[T.9]	4.846e+05	2.72e+04	17.848	0.000	4.31e+05	5.38e+05

FIGURE 9 – Résumé du modèle GLM

	Police 1	Police 2	Police 3	Police 4	Police 5
Erreur(%Ultime)	$0,0868\%$	$0,6469\%$	$0,0976\%$	$0,9694\%$	$1,4955\%$

TABLE 2 – Précision du modèle GLM sur d'autres polices

Les polices mentionnées dans la Table 2 sont de différentes tailles, ayant des différents montants de sinistralités globales. Certains triangles sont contient des cases vides, ce qui peut affecter nos estimations : Notamment, les polices 4 et 5.

FIGURE 10 – Distribution des deux clusters (K-means) par rapport aux différentes variables

Bibliographie

[1] Guillaume Beneteau. Modèle de provisionnement sur données détaillées en assurance non-vie. *ENSAE Paristech*, 2004.

[2] Karim Chayata. *La prise en charge des dépenses de santé par la solidarité nationale : l'exemple du système tunisien d'assurance maladie*. PhD thesis, Rennes 1, 2013.

[3] Equipe de l'ASJP. Estimation des provisions techniques par les méthodes déterministes et les modèles stochastiques : cas empirique du log normal. *Algerian Scientific Journal Platform*, 2020. https ://www.asjp.cerist.dz/en/downArticle/240/14/2/32248, consulté le 02/03/2023.

[4] Insurance Europe. Insurance fraud–not a victimless crime. *IE report*, 2019.

[5] Pierre-Louis Gonzalez. L'analyse en composantes principales (acp). *Tir de*.

[6] Safa Ismaïl and Nejia Zaouali. Santé et assurance maladie en tunisie : les enjeux de la réforme de 2004. *Eastern Mediterranean Health Journal*, 28(6) :444–453, 2022.

[7] Mohamed Khordj, Adlane Haffar, and Frédéric Teulon. Provisionnement et mesure de risque en assurance dommage dans le cadre de solvabilité ii. *Gestion 2000*, 34(3) :137–168, 2017.

[8] Donghyun Kim, Gian Antariksa, Melia Putri Handayani, Sangbong Lee, and Jihwan Lee. Explainable anomaly detection framework for maritime main engine sensor data. *Sensors*, 21(15) :5200, 2021.

[9] Jerzy Kowalski, Bartosz Krawczyk, and Michał Woźniak. Fault diagnosis of marine 4-stroke diesel engines using a one-vs-one extreme learning ensemble. *Engineering Applications of Artificial Intelligence*, 57 :134–141, 2017.

[10] Youguo Li and Haiyan Wu. A clustering method based on k-means algorithm. *Physics Procedia*, 25 :1104–1109, 2012.

[11] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1) :1–39, 2012.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[13] Thomas Mack. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin : The Journal of the IAA*, 23(2) :213–225, 1993.

[14] Christian Partrat. *Provisionnement technique en assurance non-vie : Perspectives actuarielles modernes*. Economica, 2007.

[15] David-Alexandre NACMIAS R. IBRAHIM. *Du provisionnement à l'évaluation du SCR pour un portefeuille de réassurance non-vie en run-off*. Institut des actuaires, 2016.

[16] Peter J Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65, 1987.

[17] Bernhard Schölkopf, Christopher JC Burges, Alexander J Smola, et al. *Advances in kernel methods : support vector learning*. MIT press, 1999.

[18] LS Shapley. A value for n-person games. RAND Paper p-295. *Santa Monica, CA : RAND*, 1952.

[19] Przemyslaw Sloma. Mémoire présenté devant l'institut du risk management pour la validation du cursus à la formation d'actuaire de l'institut du risk management et l'admission à l'institut des actuaires le 16 mai 2018. 2018.

[20] John Smith and Jane Johnson. The law of large numbers in insurance : an empirical examination. *Journal of Risk and Insurance*, 77(3) :589–604, 2010.

[21] Pierre Véron. Essai d'une statistique des fraudes commises au préjudice des compagnies d'assurances. *Journal de la société française de statistique*, 72 :49–60, 1931.