Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 6 Septembre 2023

Par : Yelly COULIBALY
Titre : Apport de la data science dans l'analyse des déclarations d'arrêts de travail en prévoyance individuelle

Confidentialité : Oui 2 ans

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

Membre présent du jury de l'Institut des Actuaires :
Romain LAÏLY
Signature :
Arthur MAILLART
Signature :
Saman NEJAD
Signature :

Entreprise :
AXA France
Signature :

Membres présents du jury de l'EURIA :
Jérémy ROUOT
Signature :

Directeur de mémoire en entreprise :
Caroline NGUY
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

L'arrêt de travail constitue un risque qui représente un coût important pour les compagnies d'assurance en France, en assurance de personnes en général, et en particulier en prévoyance individuelle. En effet, le nombre de jours en arrêt de travail est en constante augmentation depuis des années, entrainant ainsi par la même occasion la hausse des dépenses occasionnées par ce risque.
C'est donc naturellement que depuis plusieurs années, les assureurs sont à la recherche de méthodes de plus en plus précises pour analyser de manière plus efficace les demandes d'indemnisation en arrêt de travail.
Pour répondre à cet enjeu, il est important de mettre en place des études rétrospectives afin de pouvoir analyser les demandes d'indemnisations passées pour pouvoir mieux comprendre les raisons derrière l'acceptation ou le refus d'une demande plutôt qu'une autre.
Dans cette étude, nous utilisons des méthodes de machine Learning telles que les réseaux de neurones ou encore les forêts aléatoires sur des données historiques internes pour prédire l'acceptation ou le refus d'une demande d'arrêt de travail en prévoyance individuelle afin d'accompagner les gestionnaires dans l'analyse des nouvelles déclarations et de les aider dans leur prise de décision. On effectuera également des comparaisons entre ces méthodes et les modèles plus statistiques plus utilisés par les assureurs tels que le modèle linéaire généralisé (GLM) afin d'en vérifier la pertinence.
L'objectif principal de ce mémoire et donc la problématique de ce mémoire ici est de pouvoir déterminer les facteurs déterminants dans l'acceptation ou le refus d'un dossier d'ouverture d'arrêt de travail afin de réaliser une sorte de profilage des dossiers pour arrêt de travail et aider les gestionnaires dans leur prise de décision concernant ces dits dossiers.

Pour des raisons de confidentialité, certains noms dans ce mémoire seront anonymisés et les chiffres faussés.

Mots clefs: Assurance de personnes, Prévoyance individuelle, déclaration d'arrêt de travail, Data science, machine learning, Réseaux de neurones, Forêt aléatoire, GLM

Abstract

Work stoppage is a high-frequency risk that represents a significant cost for French personal insurance companies in general, and individual provident insurance in particular. In fact, the number of days of sick leave has been rising steadily over the years, driving up the cost of this risk.
So it is only natural that, for several years now, insurers have been looking for increasingly precise methods to analyze claims for lost time more efficiently.
To meet this challenge, it is important to set up retrospective studies to analyze past claims, to better understand the reasons behind the acceptance or rejection of one claim rather than another.
In this study, we are using machine learning methods such as neural networks or random forests on internal historical data to predict the acceptance or refusal of a claim for sick leave, to help managers analyze new claims and assist them in their decision-making. Comparisons will also be made between these methods and the more statistical models more commonly used by insurers, such as the generalized linear model (GLM), to verify their relevance.
The main objective of this dissertation, and therefore the problematic of this dissertation, is to be able to determine the determining factors in the acceptance or refusal of a file to open a work stoppage to carry out a kind of profiling of work stoppage files and help managers in their decision-making concerning these files.

For reasons of confidentiality, some of the names in this brief will be anonymized and the figures distorted.

Keywords: Personal insurance, Personal protection, Work stoppage declaration, Data science, machine learning, Neural networks, Random forest, GLM model

Note de Synthèse

L'arrêt de travail représente un risque qui engendre des coûts considérables pour les compagnies d'assurance en France, surtout dans le domaine de l'assurance de personnes et plus particulièrement dans la prévoyance individuelle. Au fil des années, le nombre de jours d'arrêt de travail a régulièrement augmenté, entraînant ainsi une hausse notable des dépenses liées à ce risque.

C'est pourquoi depuis un certain temps, la plupart des compagnies d'assurance dont AXA France s'est attelée à rechercher des méthodes plus précises afin d'analyser de manière plus efficace les demandes d'indemnisation pour arrêts de travail.

Afin de résoudre cette problématique, nous avons dû suivre différentes étapes qui sont les suivantes : Tout d'abord, nous avons dû constituer notre base de données.
Pour cela, nous avons procédé comme suit :

- Dans un premier temps, il a fallu procéder à l'exploration d'une nouvelle source de données, les données CRM [1]. Pour pouvoir résoudre notre problématique, il était indispensable d'avoir le maximum d'informations possibles sur les dossiers acceptés et refusés (informations sur le sinistre, le contrat associé au sinistre,...).

  Cependant, nous n'avions pas les informations relatives au sinistre pour les dossiers refusés dans les données VANDA [2] qui sont celles qui étaient le plus souvent utilisées. En effet, ces données sont très propres et contiennent les informations relatives au sinistre et au contrat pour tous les dossiers acceptés, mais elles ne contiennent pas les informations relatives au sinistre (cause du sinistre, date de survenance, date de début d'arrêt,...) pour les dossiers refusés.

  Cela rendait donc impossible toute tentative de mise en place de profil de nos dossiers. C'est la raison pour laquelle nous nous sommes tournés vers cette autre source de données qui est CRM et plus précisément vers les commentaires laissés par les gestionnaires sur cette plateforme.

1. Customer Relationship Management
2. nom anonymisé pour cause de confidentialité

Celle-ci, étant le point de départ de toute nouvelle déclaration d'arrêt de travail, nous étions à peu près sûrs d'y retrouver les informations dont nous avions besoin aussi bien pour les dossiers acceptés que refusés et c'est ce que nous avons constaté.

- Les commentaires de CRM étant très peu structurés (car écrit par différents gestionnaires avec différentes méthodes), Il nous a fallu utiliser des méthodes de text mining afin de créer de nouvelles variables.
  Ce sont par exemple les dates de début et de fin de l'AI[3], les dates de début et de fin d'hospitalisation (dans le cas où l'assuré est hospitalisé), les informations relatives aux garanties telles que les franchises accident, maladie et hospitalisation, les montants d'indemnités journalières garantis ou encore les durées maximales d'indemnisation.
- Une fois ces variables créées, nous avons dû procéder à la vérification de la concordance entre les valeurs obtenues pour les dossiers acceptés et les valeurs que nous avions dans les données VANDA.

  Le but ici était de vérifier l'efficacité de nos méthodes sur les dossiers acceptés afin d'avoir une idée de la qualité des informations récupérées pour les dossiers refusés, puis, de créer des variables finales à partir des informations obtenues dans VANDA et celles obtenues dans CRM.

  Nous avons ensuite dû compléter les valeurs manquantes de notre base, mais cela n'a rien donné de concluant et a plutôt créé un biais dans nos données.
- Nous nous sommes donc restreints à nos lignes complètes et, il a ensuite fallu réaliser une analyse descriptive de nos données afin de mieux les comprendre.
- Ensuite, nous avons mis en place nos modèles qui sont les suivants (après suréchantillonnage de la classe minoritaire "dossiers refusés") :
  - La régression logistique :
    Ce modèle a été choisi en raison de sa capacité à prédire des résultats binaires, son adaptabilité, la facilité d'interprétation de ses résultats. De plus, le fait qu'il permettent d'obtenir les odds ratio de ses variables peut être intéressant lorsque nous souhaitons mettre en place un profil.

    Les performances de ce modèle étaient "moyennes" aussi bien sur les bases d'apprentissages que sur les bases de test. En effet, nous obtenions les résultats suivants :

3. Arrêt de travail Initial

	Echantillon d'apprentissage	Echantillon de test
VN	11 060	1 890
VP	10 428	7 035
FN	5 372	3 465
FP	4 740	810
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Spécificité	0,66	0,67
Sensibilité	0,7	0,7
AUC	0,64	0,64

TABLE 1 – Caractéristique de la deuxième régression logistique sur les données d'apprentissage et de test

Cela nous a montré que le modèle ne prédisait pas très bien les dossiers refusés ; sur 100 dossiers refusés il arrivait à en prédire correctement 67. Cela n'est pas mauvais, mais nous souhaitons améliorer nos prédictions pour les dossiers refusés car prendre en charge des dossiers alors qu'on ne le devrait pas pourrait favoriser un risque de dérive du S/C.

— La forêt aléatoire :
Nous avons choisi ce modèle pour sa précision dans les prédictions, son évolutivité (il peut être utilisé sur efficacement sur de grands volumes de données) et sa robustesse (il est moins sensible au risque de surapprentissage).
C'est ce modèle qui nous donne les meilleures performances, nous pouvons les observer dans le tableau ci-dessous :

	Échantillon d'apprentissage	Échantillon de test
VN	15 642	2 160
VP	15 484	8 190
FN	316	2 310
FP	158	540
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Spécificité	0,98	0,78
Sensibilité	0,99	0,8
AUC	0,99	0,89

TABLE 2 – Caractéristique du modèle de forêt aléatoire sur les données d'apprentissage et de test

Ces résultats montrent que ce modèle arrive à plutôt bien prédire les dossiers refusés, en effet, sur 100 dossiers refusés il permet d'en prédire 78 correctement.
Cela n'est pas parfait mais est déjà beaucoup plus satisfaisant que ce qu'on obtient avec la régression logistique.

— Le réseau de neurones :
Nous avons utilisé ce modèle car il peut gérer des données complexes et capturer des relations subtiles entre les variables mais également pour la précision de ses prédictions.
Avec ce modèle, nous obtenons les résultats suivants :

	Echantillon d'apprentissage	Échantillon de test
VN	14 536	2 187
VP	14 062	7 665
FN	1 738	2 835
FP	1 264	513
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Spécificité	0,89	0,73
Sensibilité	0,92	0,81
AUC	0,91	0,82

TABLE 3 – Caractéristiques du réseau de neurones sur les données d'apprentissage après sur-échantillonage et les données de test

Ces résultats nous permettent de voir que ce modèle est aussi plus efficace que la régression logistique pour prédire correctement les dossiers refusés, sur 100 dossiers refusés, il arrive à en prédire correctement 73.

— Et enfin, la comparaison de nos différents modèles nous a permis de nous rendre compte que le plus adapté à nos données est le modèle de forêt aléatoire car il prédit mieux les dossiers refusés que les autres.

Executive summary

Work stoppage is a risk that generates considerable costs for insurance companies in France, especially in the personal insurance sector. Over the years, the number of days of sick leave has risen steadily, resulting in a significant increase in expenses linked to this risk.

For this reason, most insurance companies, including AXA France, have been working for some time to find more accurate methods of analyzing work stoppage claims more efficiently.

In order to solve this problem, we had to follow the following steps :
First, we had to build up our database. To do this, we proceeded as follows :

- First, we had to explore a new data source, CRM data [4]. In order to solve our problem, it was essential to have as much information as possible on accepted and rejected files (information on the claim, the contract associated with the claim, ...). However, we did not have the claims information for rejected files in the VANDA data [5], which is the one most often used. In fact, this data is very clean and contains information relating to the claim and the contract for all accepted files, but it does not contain information relating to the claim (cause of claim, date of occurrence, date of start of stoppage, etc.) for refused files.

  This made any attempt to profile our files impossible. This is why we turned to another data source, CRM, and more specifically to the comments left by managers on this platform.

  As this is the starting point for all new work stoppage declarations, we were pretty sure that it would contain the information we needed for both accepted and rejected cases, and this is what we found.

- As the CRM comments are not very structured (because they were written by different managers using different methods), we had to use text mining methods to create new variables. These include, for example, the start and end dates of the IWS [6], the start and end dates of hospitalization (if the insured is hospitalized), infor-

4. Customer Relationship Management
5. name anonymized for reasons of confidentiality
6. Initial Work Stoppage

mation relating to benefits such as accident, illness and hospitalization deductibles, guaranteed daily benefit amounts and maximum benefit periods.

- Once these variables had been created, we had to check that the values obtained for the accepted files matched the values we had in the VANDA data.

  The aim here was to check the effectiveness of our methods on accepted files, in order to get an idea of the quality of the information retrieved for refused files, and then to create final variables from the information obtained in VANDA and that obtained in CRM.

  We then had to fill in the missing values in our database, but this didn't produce any conclusive results, and instead created a bias in our data.
- We therefore restricted ourselves to our complete lines, and then had to carry out a descriptive analysis of our data in order to better understand them.
- Next, we set up our models, which are as follows (after oversampling the minority class "refused files") :
  - Logistic regression :
    This model was chosen for its ability to predict binary outcomes, its adaptability, its ease of interpretation and its results. For example, it allows us to obtain odds ratios for its variables, which can be useful when we want to set up a profile.

    The performance of this model was average on both the learning and test bases. Indeed, we obtained the following results :

	Learning sample	Test sample
VN	11 060	1 890
VP	10 428	7 035
FN	5 372	3 465
FP	4 740	810
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Specificity	0.66	0.67
Sensitivity	0.7	0.7
AUC	0.64	0.64

TABLE 4 – Characteristic of second logistic regression on training and test data

    This showed us that the model didn't predict refused cases very well ; out of 100 refused cases, it managed to correctly predict 67. This is not bad, but we would like to improve our predictions for refused cases, as taking on cases when we shouldn't could lead to a risk of S/C drift.
  - The random forest :
    We chose this model for its accuracy in predictions, its scalability (as it can be used efficiently on large volumes of data) and its robustness (it is less sensitive

to the risk of overlearning).
This is the model that gives us the best performance, as shown in the table below :

	Learning sample	Test sample
VN	15 642	2 160
VP	15 484	8 190
FN	316	2 310
FP	158	540
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Specificity	0.98	0.78
Sensitivity	0.99	0.8
AUC	0.99	0.89

TABLE 5 – Characteristic of random forest model on training and test data

Our results show that this model does a pretty good job of predicting rejected files : out of 100 rejected files, it predicts 78 correctly. This is not perfect, but it is already much more satisfactory than what we get with logistic regression.
neural network :
We used this model because it can handle complex data and capture subtle relationships between variables, but also because of its accuracy.
With this model, we obtain the following results :

	Learning sample	Test sample
VN	14 536	2 187
VP	14 062	7 665
FN	1 738	2 835
FP	1 264	513
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Specificity	0.89	0.73
Sensitivity	0.92	0.81
AUC	0.91	0.82

TABLE 6 – Features of the neural network on training data after oversampling and test data

These results show that this model is also more effective than logistic regression in correctly predicting rejected files : out of 100 rejected files, it correctly predicts 73.

— And finally, comparing our different models, we realized that the best suited to our data is the random forest model, as it predicts refused files better than the others.

Remerciements

Mes remerciements vont à l'endroit de :

L'administration et du corps professoral de l'Euro Institut d'Actuariat (EURIA) avec à sa tête M. VERMET, pour ces années d'apprentissage et de formation dans le domaine de l'actuariat.

Mme Anaëlle LEBERRE, ma tutrice de mémoire pour sa disponibilité, son encadrement pédagogique et son soutien tout au long du stage.

M. Brice FRANKE, mon tuteur académique, pour ses encouragements et son soutien.

AXA pour l'opportunité et la confiance accordée tout au long de cette alternance.

Mme Caroline NGUY, ma tutrice chez AXA France ainsi que M. Guillaume ROUSTIDE, le manageur de l'équipe Predictive et data, pour leur disponibilité, leurs explications et leurs conseils.

Ma mère et mon père, pour les efforts et leur amour inconditionnel et sans qui ce travail n'aurait pas vu le jour ainsi qu'à tous les membres de ma famille.

Mes ami.e.s pour l'amitié et leurs contributions à ce travail.

Tous ceux qui de près ou de loin m'ont soutenue et ont contribué à l'aboutissement de ce travail, que chacun trouve ici l'expression de ma profonde gratitude.

Table des matières

Executive summary **ix**

1 Contexte **5**
1.1 L'assurance de personnes 6
1.1.1 Généralités 6
1.1.2 La prévoyance individuelle : définition 7
1.1.3 Les garanties en prévoyance individuelle 7
1.1.4 La sinistralité en prévoyance individuelle, pour la garantie incapacité temporaire de travail 10
1.2 La data science : définition, origine, importance et application 12
1.2.1 Définition de la data science 12
1.2.2 Origine de la data science 12
1.2.3 Importance et application de la data science 13
1.2.4 Intégrité éthique des données de recherche 14

2 Modèles de machine learning **17**
2.1 Généralités 18
2.2 Modèle linéaire généralisé 18
2.2.1 Théorie 18
2.2.2 Modèle linéaire pour les variables binomiales : régression logistique 19
2.2.3 Amélioration des performances de la régression logistique 22
2.3 Forêt aléatoire 26
2.3.1 Théorie 26
2.3.2 Méthodes d'amélioration des performances du modèle 27
2.3.3 Importance des variables 28
2.4 Réseaux de neurones 29
2.4.1 Théorie 29
2.4.2 Rétropropagation 29
2.4.3 Importance des variables ou caractéristiques 30

3 Données et traitement des valeurs manquantes **31**
3.1 Présentation des données . . . 32
3.2 Utilisation des méthodes de Text Mining pour la création de nouvelles variables . . . 33
3.2.1 Présentation du text mining . . . 33
3.2.2 Jointure des tables CRM importantes . . . 34
3.2.3 Création de nouvelles variables à partir des commentaires de la base CRM . . . 38
3.3 Vérification de la conformité des méthodes utilisées . . . 44
3.3.1 Ajout des données VANDA . . . 44
3.3.2 Tests de conformité et création des variables finales . . . 47
3.4 Vérification de la complétude des variables finales . . . 52
3.5 Traitement des valeurs manquantes pour les variables finales . . . 57
3.5.1 Mise en place de modèles de machine learning et sélection des meilleurs modèles . . . 57
3.5.2 Récapitulatif des variables complétées et des méthodes utlisées . . . 60
3.6 Vérification de la cohérence de nos prédictions . . . 62

4 Analyse des données **65**
4.1 Analyse descriptive des données . . . 66
4.1.1 Analyse descriptive univariée . . . 66
4.1.2 Analyse descriptive bivariée . . . 72
4.1.3 Analyse par rapport à l'OUTCOME . . . 77
4.1.4 Apport des résultats des analyses . . . 80
4.2 Étude de corrélation entre les variables . . . 80
4.2.1 Corrélation de Pearson . . . 81
4.2.2 Test de Khi 2 . . . 85
4.3 Variables à conserver pour nos modèles . . . 87

5 Modélisation : prédiction de la décision pour un nouvel arrêt de travail **89**
5.1 Division de notre base de données . . . 90
5.2 Régression logistique . . . 90
5.2.1 Premier modèle de régression logistique : le modèle complet . . . 90
5.2.2 Sélection des variables significatives et obtention du modèle de régression logistique optimisé . . . 91
5.2.3 Suréchantillonage de la classe minoritaire (dossiers refusés) . . . 94
5.2.4 Prise en compte de l'interaction entre certaines variables . . . 95
5.2.5 Interprétation des résultats . . . 95
5.3 Forêt aléatoire . . . 97

5.3.1 Application d'un modèle de forêt aléatoire à nos données 97
5.3.2 performances du modèle sur les données après suréchantillonnage de la classe minoritaire 99
5.3.3 Analyse et interprétation des résultats 101
5.4 Réseau de neurones 102
5.4.1 Application du réseau de neurones à nos données 102
5.4.2 Performances du modèle après suréchantillonnage de la classe minoritaire 103
5.4.3 Analyse des résultats 104
5.5 Comparaison des différents modèles et choix du plus adapté 105

A Étapes de la jointure des données CRM aux données VANDA **113**

B Résultats complets du test de Khi 2 **117**

C Odds Ratios **121**

D Importance des variables avec le réseau de neurones **123**

Bibliographie **130**

Glossaire

ACC = Accident
AI = Arrêt de travail Initial
AT = Arrêt de Travail
AV = Avant
BH = Bulletin d'Hospitalisation
CRH = Compte Rendu d'Hospitalisation
CRM = Customer Relationship Management
DECLA = Déclaration
GAR = Garantie
HOSPI = Hospitalisation
IJ = Indémnités Journalières
MAL = Maladie
NSM = Niveau de Sélection Médicale
RO = Régime Obligatoire
SURV = Survenance
VANDA = Nom anonymisé d'une source de données
VF = Version Finale

Introduction

La data science a su en peu de temps gagner en popularité et devenir quasiment indispensable, et ce, dans pratiquement tous les domaines : la santé pour faciliter l'analyse des dossiers médicaux, la finance pour l'analyse des données des marchés financiers, les sciences sociales pour l'analyse des données démographiques ou encore l'assurance pour l'amélioration de la tarification des polices par exemple.

Le secteur de l'assurance, en particulier celui de l'assurance de personne dont l'activité repose essentiellement sur les données pour adapter les offres et prévoir les sinistres entre autres, s'est vu beaucoup impacté par cette science au cours de ces dernières années. En effet, on se sert de la data science aujourd'hui en assurance de personne pour détecter les fraudes, évaluer de manière plus précise les risques et même parfois pour prendre des décisions plus éclairées et basées sur les données historiques.

La data science comme moyen de prendre des décisions plus éclairées reste peu utilisée en prévoyance individuelle, surtout pour l'analyse des nouvelles déclarations d'arrêts de travail. Néanmoins, les assureurs s'y intéressent de plus en plus, car ils sont en constante recherche de nouvelles méthodes et outils innovants pour faciliter l'analyse de ce risque et améliorer sa connaissance.

Les travaux de ce mémoire se focaliseront sur l'application pratique de la data science, plus précisément des méthodes de machine learning, pour prédire la décision pour une nouvelle déclaration d'arrêt de travail reçue par un assureur en prévoyance individuelle. Nous pourrons alors nous baser sur les facteurs prédictifs obtenus pour en faciliter l'analyse et la prise de décision de ses gestionnaires.

La démarche entreprise dans ce mémoire consistera à utiliser les données fournies par l'assureur pour segmenter le portefeuille grâce à des algorithmes qu'on n'utilise pas couramment en prévoyance individuelle. L'objectif étant d'identifier les facteurs qui peuvent prédire l'acceptation ou le rejet des nouvelles demandes d'indemnisation en arrêt de travail.

Chapitre 1

Contexte

1.1 L'assurance de personnes

1.1.1 Généralités

L'assurance de personnes est un service qui permet de fournir une prestation à des personnes physiques lorsque survient un évènement **aléatoire** tel qu'une maladie, un accident corporel, la dépendance, l'invalidité ou encore le décès. Contrairement à l'assurance **IARD** qui couvre seulement les dommages matériels, l'assurance de personnes, elle se concentre sur la protection de l'individu lui-même et des membres de sa famille s'ils sont déclarés comme bénéficiaires.
Elle regroupe entre autres l'épargne, la retraite, la santé et la prévoyance et selon le mode de souscription, on en distingue deux types : **la collective** et **l'individuelle** (**confère figure 1.1**).

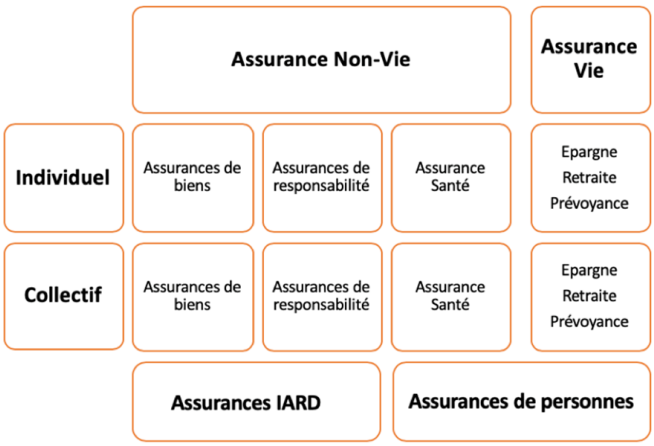


FIGURE 1.1 – Les différentes catégories des produits d'assurance (Source :[13])

Lorsqu'elle est souscrite de manière collective, on parle d'**assurance de personne collective**. Celle-ci est la plupart du temps souscrite par un employeur ou une autre entité de groupe qui peut être par exemple une association professionnelle ou un syndicat et a pour but de fournir une couverture d'assurance à un ensemble de personnes qui doivent avoir un lien de même nature avec le souscripteur et répondre à des conditions définies au contrat.
Ce lien peut être :

— Le lien entre un salarié et son employeur, dans ce cas, c'est l'employeur qui se charge de la souscription au contrat,

— Le lien entre une organisation syndicale et ses membres.

De plus, elle représente des avantages pour l'assuré, car elle permet :

— Une mutualisation des risques,

— Des niveaux de remboursements généralement plus élevés que pour une assurance individuelle (surtout en santé),
— Des cotisations moins élevées entre autres, car les primes sont souvent partagées entre l'employeur et les employés avec une partie déduite des salaires ou des avantages sociaux des employés.

Lorsqu'elle est souscrite de manière individuelle, on parle d'**assurance de personne individuelle**. Ce type d'assurance est souscrit directement par l'individu assuré, il paie lui-même ses primes directement à l'assureur pour bénéficier de la couverture d'assurance.
Elle permet donc d'avoir un produit personnalisé en fonction de besoins spécifiques à chaque individu et donc plus ou moins avantageux pour les assurés. En effet, ses primes par exemple sont basés sur des facteurs beaucoup plus précis que ceux de l'assurance collective (âge, état de santé, couverture désirée ...). Nous nous intéresserons particulièrement à ce type d'assurance de personnes et plus précisément à la prévoyance individuelle dans la suite de ce mémoire.

1.1.2 La prévoyance individuelle : définition

La prévoyance individuelle est une branche de l'assurance de personne individuelle. Elle est souscrite individuellement par chaque assuré et leur offre une protection financière en cas de survenance d'évènements **aléatoires** qui les empêchent en partie ou totalement de continuer à exercer leurs activités professionnelles, entraînant ainsi une perte de revenu ou la perte d'autonomie financière.
Elle regroupe entre autres les garanties suivantes :

— La garantie **invalidité**,
— La garantie **décès**,
— La garantie **dépendance**,
— La garantie **obsèques** ou encore
— La garantie **incapacité temporaire de travail**.

Et chaque garantie a une manière de fonctionner et un but bien précis que nous énumérerons dans la suite de notre travail.

1.1.3 Les garanties en prévoyance individuelle

i) La garantie invalidité

L'invalidité est un état dans lequel un individu perd sa capacité à travailler au moins à 66% au sens de la sécurité sociale en France. Le degré de l'état d'invalidité est déterminé à partir d'un taux appelé **taux d'invalidité,** lui-même déterminé en fonction d'un barème proposé par la compagnie d'assurance, ce barème peut être différent de celui établit par la sécurité sociale.
Pour un arrêt de travail provoqué par une maladie ou un accident de la vie personnelle, la sécurité sociale a défini 3 types d'invalidité :

— La **première catégorie** : elle regroupe les personnes invalides capables d'exercer une activité rémunérée,

— La **deuxième catégorie** : elle regroupe les personnes invalides absolument incapables d'exercer une activité professionnelle quelconque et enfin,

— La **troisième catégorie** : elle regroupe les personnes invalides incapables d'exercer une profession et dans l'obligation de recourir à l'assistance d'une tierce personne pour effectuer les actes ordinaires de la vie.

Dans le cas d'un accident du travail ou d'une maladie professionnelle, le taux d'incapacité permanente dépend de la nature de l'infirmité, l'âge, les facultés mentales ou physiques de l'assuré, entre autres [21]. La garantie invalidité prévoit le versement d'un montant au bénéficiaire désigné de la police d'assurance en cas d'invalidité totale ou partielle qui l'empêche de continuer à exercer son activité professionnelle, et ce, en fonction du degré de l'invalidité.

L'assureur prévoit parfois des clauses qui dans certaines situations excluent toute forme de prise en charge du sinistre. Cela peut être :

— La pratique d'un sport à risque,

— Des situations de guerre ou encore

— Des attaques terroristes...

Le montant versé peut-être soit un capital payé une seule fois par l'assureur ou une rente versée de manière fractionnée par année, elle peut être semestrielle, trimestrielle ou mensuelle par exemple.

ii) La garantie décès

C'est une garantie qui prévoit le versement d'un certain montant à des bénéficiaires choisis par l'assuré, ce montant peut-être, comme dans le cas de la garantie en cas d'invalidité, un capital ou une rente.
Le but de cette garantie est d'apporter une protection financière aux bénéficiaires qui sont en général la famille et les proches de l'assuré.

Elle peut aussi comprendre des clauses d'exclusions :

— Le décès dû à une maladie survenue avant la souscription du contrat,

— Ou encore le suicide au cours de la première année après la souscription du contrat.

iii) La garantie dépendance

Pour comprendre cette garantie, il faut d'abord comprendre le terme **dépendance**.
Ce terme en assurance fait référence à une situation dans laquelle une personne a besoin

d'une autre de manière permanente et irréversible pour effectuer les actes de la vie quotidienne tels que se laver, se déplacer ou encore s'alimenter. C'est le résultat du vieillissement et de maladies chroniques telles que Parkinson, les maladies cardiovasculaires ou encore Alzheimer en général, et rarement le résultat d'un accident.
Pour palier à ce problème, l'assurance prévoit le versement d'une rente (qui est la plupart du temps mensuelle) ou d'un capital pour couvrir les frais liés aux soins et à l'assistance pour l'assuré en état de dépendance.
Elle peut, elle aussi, avoir des clauses d'exclusions dont les plus courantes sont :

— Un accident de la route, avec consommation de boissons alcoolisées constatée par une alcoolémie égale ou supérieure au taux autorisé par la législation en vigueur au moment du sinistre,

— L'usage de stupéfiants non prescrits pour un usage médical

— Une guerre civile ou étrangère, une émeute, une rixe ou des actes de terrorisme dans lesquels l'assuré a pris une part active

— Des complications physiques ou neuropsychiatriques dues à l'abus chronique de consommation de boissons alcoolisées

— Des risques aériens se rapportant à des compétitions organisées dans un cadre officiel ou privé, des tentatives de records, des vols d'essai sur prototype ...

iv) La garantie obsèques

La garantie obsèques, elle, permet de couvrir les dépenses liées aux obsèques de l'assuré après son décès et ainsi de protéger ses proches des dépenses engendrées. Ainsi, le souscripteur au moment de son décès s'assure que les fonds sont versés directement à l'organisateur des funérailles afin que ceux-ci puissent être organisés selon sa volonté. Le capital est à constituer par versement de primes uniques, temporaires ou viagères et peut atteindre jusqu'à 10 000 € et permet de financer :

— Le cercueil,

— La cérémonie funéraire,

— La préparation du corps ...

Et certains contrats permettent même de financer :

— Le déplacement des proches jusqu'au lieu des obsèques,

— Une assistance psychologique pour les proches du défunt,

— Le rapatriement de la dépouille si cela s'avère nécessaire ...

Elle comprend en général les mêmes clauses d'exclusions que la garantie décès.

v) La garantie incapacité temporaire de travail

Cette garantie permet d'offrir une protection financière à l'assuré en cas de survenance d'une maladie ou d'un accident le laissant dans l'incapacité temporaire de travailler. Elle prévoit le paiement d'indemnités journalières au bénéficiaire pour l'aider à compenser la perte de ses revenus d'activité pendant la période d'incapacité et donc à maintenir son niveau de vie.
Elle peut être reconduite plusieurs fois si cela s'avère nécessaire pour une durée totale de 3 ans maximum au-delà desquels l'assuré est déclaré en **invalidité** si son état ne s'est toujours pas amélioré et qu'il est toujours dans l'incapacité de reprendre son activité professionnelle.

Comme la garantie invalidité, elle peut, elle aussi, contenir des clauses d'exclusions de prise en charge du sinistre dans certains cas (en général, les mêmes que ceux de la garantie invalidité).
Dans la suite de notre étude, nous nous intéresserons particulièrement à cette garantie dans le cadre des arrêts de travail.

1.1.4 La sinistralité en prévoyance individuelle, pour la garantie incapacité temporaire de travail

Le risque arrêt de travail est un risque qui peut s'avérer compliqué à étudier compte tenu du fait qu'il touche à différents domaines en même temps tels que la santé, l'économie et le social.

Ces dernières années, les dépenses engendrées par ce risque n'ont cessé d'augmenter du fait de la hausse de l'absentéisme au travail partout en France, ce qui a pesé lourdement sur les collectivités, les entreprises et les assurances.
En effet, le taux d'absentéisme qui avait connu une hausse d'environ 3% entre 2016 et 2017, en a connu une nouvelle de 8% en 2018, ce qui a provoqué l'atteinte d'un taux global de 05,10%, représentant ainsi une moyenne de 18,6 jours d'absence pour un salarié [23].

Plus récemment, entre 2019 et 2021, pendant la période de crise sanitaire liée à la Covid-19, ce taux a encore augmenté du fait des cas de covid, mais aussi des demandes d'arrêt pour mal de dos et a même atteint un pic de 06,7% en 2022.
Ainsi, entre 2017 et 2022, ce taux a connu une hausse de **37%** et une hausse de **15%** entre 2019 et 2022 [23].

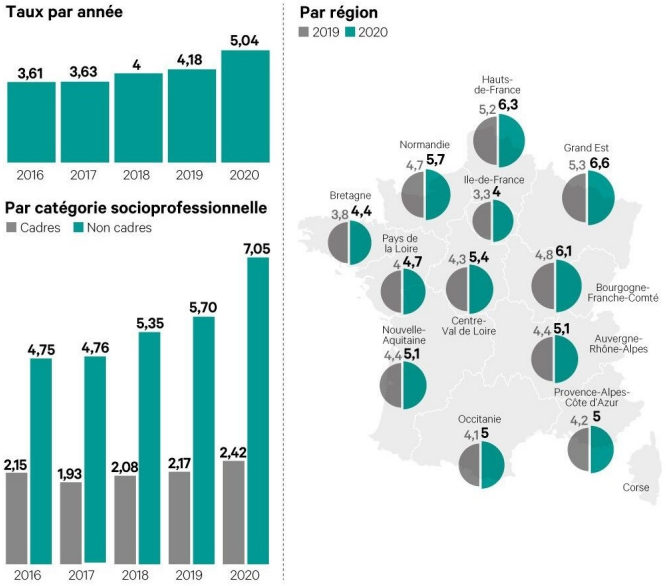


FIGURE 1.2 – L'absentéisme aggravé par la crise de la Covid en France (Source : [18])

Cette hausse quasiment constante de l'absentéisme liée à ce risque a replacé la sinistralité comme l'un des enjeux majeurs chez les assureurs, les poussant ainsi à se tourner vers la **data science** afin d'explorer d'autres moyens pour pouvoir prévoir leur sinistralité.

1.2 La data science : définition, origine, importance et application

1.2.1 Définition de la data science

La data science est une discipline qui permet d'analyser, de comprendre et de donner un sens aux données. Elle regroupe l'usage de la technologie informatique et des statistiques pour structurer les données afin qu'elles puissent être facilement comprises, une fois les données analysées, les data scientists sont en mesure de conseiller et de mieux orienter les décisions de l'entreprise.

C'est une branche pluridisciplinaire qui allie connaissances en mathématiques et statistiques, outils informatiques, algorithmiques et une bonne connaissance du monde de l'entreprise. L'un de ses objectifs principaux étant d'aider les entreprises à faire des choix stratégiques, leur permettant ainsi d'augmenter leur chiffre d'affaires.
Les entreprises dans pratiquement tous les secteurs d'activité, qu'il s'agisse de l'informatique, la santé ou encore l'assurance l'utilisent systématiquement au quotidien.

1.2.2 Origine de la data science

L'histoire de la data science remonte à l'époque de la collecte et de l'analyse des données. Cependant, c'est avec l'avènement de l'informatique et l'émergence des nouvelles technologies que la data science a véritablement pris son envol.

C'est au XVIIe siècle que les premières méthodes statistiques ont été développées par des penseurs tels que Blaise Pascal [26] et Pierre Simon Laplace. Ces méthodes ont permis de poser les bases nécessaires à l'analyse des données et à la prise de décisions basée sur des observations puis, entre les années 40 et 50 avec l'émergence de l'informatique et la création des ordinateurs, il devint possible de stocker d'immenses quantités de données et c'est alors que des précurseurs tels que John W. Tukey et Norbert Wiener ont alors commencé à explorer des méthodes d'analyse statistique appliquées à ces précieuses données.

Dans les années 50 et 60, des chercheurs éminents comme Alan Turing [1] et John McCarthy [5] ont jeté les bases fondamentales du domaine qu'est aujourd'hui l'intelligence artificielle (IA). Cette avancée majeure a ouvert la voie à une branche spécifique de la data science appelée **apprentissage automatique (machine learning)**, permettant aux ordinateurs d'apprendre et de s'améliorer continuellement en utilisant les données. Et, à partir des années 90, grâce à une numérisation toujours croissante des données et l'arrivée d'Internet, la quantité des données générées a connu une montée en flèche. Cela a donc donné lieu à de nouveaux défis et à de nouvelles perspectives en matière d'analyse des données. En conséquence, il est devenu plus que nécessaire d'établir des techniques plus avancées capables d'extraire des informations significatives à partir de vastes ensembles de données.

1.2.3 Importance et application de la data science

Dans tous les secteurs d'activité

Le volume des données mondiales atteindra **175 zettaoctets** d'ici à 2025 ; en plus, la data science est utilisée dans une grande variété de domaines et secteurs d'activité [6]. C'est la raison pour laquelle cette discipline s'avère être un outil d'une importance capitale pour les entreprises de nos jours.

Il serait difficile de citer toutes ses applications ici, mais il est plus aisé d'analyser le caractère stratégique de la data science en entreprise.
Ses principales utilisations sont par exemple :

— Permettre de mesurer, suivre et optimiser les performances des entreprises en améliorant leurs produits et services, mais jour également un rôle vital dans presque tous les aspects opérationnels et commerciaux de leur activité, en d'autres termes, elle sous-tend chaque décision stratégique d'une organisation prouvant sa pertinence et son efficacité,

— Aider à une meilleure compréhension des résultats engendrés par chacune des actions de l'entreprise et apprendre des expériences passées,

— Eťant de plus en plus utilisée dans le cadre de l'analyse des fluctuations financières et donc dans la prédiction de leurs résultats sur le moyen et long terme. Les données sur les éléments du bilan d'une entreprise étant des représentants importants des indices de croissance de celle-ci, la data science permet de mieux les comprendre et de se prémunir des risques financiers (en assurance en détectant les fraudes par exemple), enfin,

— Elle permet aussi de disposer d'un large choix de méthodes permettant ainsi d'accroître l'efficacité de la gestion et la fourniture des produits ou services de l'entreprise concernée.

Tout ceci nous montre bien l'importance de cette discipline dans une entreprise.

Dans l'assurance en particulier

La data science se révèle d'autant plus importante sur le marché de l'assurance, car en plus des avantages qu'elle représente pour pratiquement tous les secteurs d'activité, elle a aussi des avantages qui sont propres au secteur de l'assurance.

Comme nous l'avons déjà expliqué, les assureurs se retrouvent à exploiter d'énormes volumes de données de nos jours. Ce qui fait que leur utilisation ne se cantonne plus seulement à l'aspect actuariel (tarifer les produits, détecter les fraudes, améliorer les techniques de provisionnement,...) mais va bien au-delà [2].

En assurance, la data science permet aux assureurs de :

— Améliorer la qualité du service proposé aux assurés : en effet, de nos jours, les compagnies d'assurance proposent des services de plus en plus adaptés aux besoins de leurs clients (ou futurs clients) ce qui permet de fidéliser la clientèle déjà acquise et en même temps d'augmenter ses parts de marché.

— Anticiper certaines perturbations naturelles majeures telles que les inondations, les tempêtes, les séismes, ... grâce aux données météorologiques et géologiques, ce qui permet d'offrir des mesures de précaution aux assurés et permet aux assureurs de proposer des garanties plus adaptées au risque encouru. Dans certains endroits, on ne peut par exemple pas être couvert pour certains risques climatiques ou on est amené à payer des primes beaucoup plus élevées si on souhaite tout de même être couvert. C'est le cas des zones inondables à haut risque par exemple. Cela permet à l'assureur de minimiser le risque d'être dans l'incapacité de tenir ses engagements [3].

1.2.4 Intégrité éthique des données de recherche

Comme pour toutes les sciences, la question de l'intégrité scientifique et éthique a dû être abordée. En effet, suite au scandale lié à l'exploitation illicite des données personnelles et qui a secoué Facebook en 2018 [8] et qui a éclaté quand la société Cambridge Analytica fut mise à nu pour sa collecte frauduleuse des renseignements de millions d'utilisateurs via une application ludique, un texte réglementaire européen encadrant le traitement des données de manière égalitaire sur tout le territoire de l'Union européenne (UE) a été mis en place [9]. Il y a eu un véritable séisme qui s'est créé en matière de confidentialité et d'éthique, jetant une lumière crue sur les failles de sécurité des données sur les réseaux sociaux.

Dans l'univers de l'assurance, cette controverse a amplifié la prise de conscience quant à la préservation des données et au secret. En effet, les compagnies d'assurance rassemblent des informations confidentielles sur les souscripteurs, par exemple pour évaluer les risques et déterminer les cotisations. La mise à jour sur la fragilité des données a incité les assureurs à fortifier leurs mesures de sécurité en vue de préserver la confidentialité des informations personnelles de leurs clients.

Cette situation a fait que certaines données sont soumises à d'importantes clause de confidentialité et seulement un nombre très limité de personnes peuvent y avoir accès. C'est par exemple le cas des informations sur les pathologies ayant provoqué un arrêt de travail en prévoyance. Ces informations peuvent être très importantes dans le cadre des analyses des arrêts de travail, mais étant des données critiques, elles ne peuvent être utilisées dans ce cadre et cela pourrait favoriser l'obtention de résultats parfois biaisés.

Dans la suite de notre étude, nous évoquerons de manière plus détaillée la branche de la data science qui nous intéresse le plus, à savoir **l'apprentissage automatique,** plus

communément appelé **Machine learning** et nous expliquerons comment cela nous a aidé à répondre à la problématique de notre étude.

Chapitre 2

Modèles de machine learning

2.1 Généralités

Le machine learning traduit par l'expression "apprentissage automatique ou automatisé" en français, se définit comme un code informatique permettant aux ordinateurs d'être capables de réaliser un apprentissage autodirigé grâce à la mise en place d'algorithmes et l'utilisation des probabilités à partir de données de départ.
La machine se met alors en apprentissage afin d'acquérir de nouvelles compétences, de réagir de manière adaptée lorsqu'elle rencontre un problème et de prendre des décisions en fonction de différentes situations plutôt que de simplement suivre des instructions à la lettre.

Cette discipline repose sur deux éléments très importants : **les données** et **l'algorithme d'apprentissage**.
Elle nécessite donc d'avoir des données propres et de bonne qualité, sinon, les prévisions obtenues risquent d'être très peu fiables et les résultats peu voir pas du tout exploitables, c'est le concept du *"garbage in garbage out"* [24].
Pour notre étude, les modèles avec lesquels nous choisissons de travailler sont les suivants :

— La régression logistique :
ce modèle a été choisi en raison de sa capacité à prédire des résultats binaires, son adaptabilité, et la facilité d'interprétation de ses résultats. De plus, le fait qu'il permette d'obtenir les odds ratio des variables explicatives peut être intéressant lorsque nous souhaitons mettre en place un profil de nos données.

— La forêt aléatoire :
Nous avons choisi ce modèle pour sa précision dans les prédictions, son évolutivité (il peut être utilisé efficacement sur de grands volumes de données) et sa robustesse (il est moins sensible au risque de surapprentissage).

— Le réseau de neurones :
Nous avons utilisé ce modèle car il peut gérer des données complexes et capturer des relations subtiles entre les variables explicatives mais également pour la précision de ses prédictions.

Dans la suite de notre mémoire, nous mettrons l'accent sur les modèles que nous avons utilisés pour compléter nos valeurs manquantes, mais aussi pour répondre en partie à notre problématique de départ, à savoir les utiliser pour analyser les arrêts de travail.

2.2 Modèle linéaire généralisé

2.2.1 Théorie

Contrairement aux modèles linéaires classiques qui reposent sur les hypothèses de continuité de la variable réponse (ce qui implique qu'il soit applicable uniquement aux variables continues) et une relation linéaire entre celle-ci et les variables explicatives, le

modèle linéaire généralisé, communément appelé GLM est un modèle adapté à des variables réponses de natures diverses.
Nous pouvons par exemple l'utiliser pour des variables de type binaire, à valeur dans $\mathbb{N}$, ou encore des variables à valeur dans $\mathbb{R}_+$ [10].
Ce modèle suppose une relation linéaire entre une fonction de la moyenne de la variable réponse appelée **fonction lien** et les variables prédictives.
Pour un jeu de données avec n lignes et m colonnes, pour lequel Nous appellerons Z_k la variable réponse et $X_1, X_2, \ldots, X_q$ les variables explicatives de notre modèle $(q \in 1, \ldots, m)$, sa structure repose sur les trois hypothèses ci-dessous :

— L'indépendance des variables aléatoires $Z_1, Z_2, \ldots, Z_n$,

— Les paramètres de la loi de Z_k dépendent des variables explicatives et

— L'existence d'une relation linéaire entre une fonction de l'espérance de la loi de la variable réponse et les variables explicatives. Celle-ci peut être représentée par l'équation suivante :

$$g(E[Z_i]) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_q x_{i,q} \tag{2.1}$$

Où :

- g est la fonction de lien qui doit avoir un domaine correspondant à l'intervalle de valeurs possibles de la variable réponse et être inversible, on note que pour la régression logistique g est la fonction *logit* définie comme suit :

$$\begin{aligned}]0,1[&\longrightarrow \mathbb{R} \\ x &\longmapsto g(x) = ln(\frac{x}{1-x}) \end{aligned}$$

- $\beta = (\beta_1, \ldots, \beta_q)$, $\beta_j \in \mathbb{R}$ les paramètres du modèle et
- $x_{i,1}, x_{i,2}, \ldots, x_{i,q}$ les réalisations des variables aléatoires $X_1, X_2, \ldots, X_q$.

Dans la suite, nous verrons comment adapter ces hypothèses à la régression logistique.

2.2.2 Modèle linéaire pour les variables binomiales : régression logistique

Fonction de vraisemblance

La régression logistique est utilisée pour expliquer une variable de type binaire, il s'agit d'une variable qualitative. La variable réponse peut par exemple prendre les valeurs "survie" et "décès", "malade" et "sain" ou encore "accepté" et "refusé" comme pour notre étude.
Nous souhaitons estimer les paramètres inconnus β du GLM binomial en utilisant la méthode maximum de vraisemblance.
La fonction de vraisemblance ici est appelée $L(\beta)$ et est calculée comme suit :

$$\begin{aligned} L(\beta) &= P[Z_1 = z_1, Z_2 = z_2, \ldots, Z_q = z_n; \beta] \\ &= \prod_{i=1}^{n} P[Z_i = z_i; \beta] \end{aligned}$$

(car les variables $(Z_i)_{i\in 1,\ldots,n}$ sont supposées indépendantes).
Nous savons que si les variables aléatoires $(Z_i)_{i\in 1,\ldots,n}$ suivent une loi binomiale :

$$P[Z_i = z_i] = r(x_{i,1}, \ldots, x_{i,q})1\{z_i = 1\} + (1 - r(x_{i,1}, \ldots, x_{i,q}))1\{z_i = 0\}$$

nous pouvons aussi l'écrire comme suit :
$$P[Z_i = z_i] = r(x_{i,1}, \ldots, x_{i,q})^{z_i}(1 - r(x_{i,1}, \ldots, x_{i,q}))^{1-z_i}$$

Avec $r(x_{i,1}, \ldots, x_{i,q})$ la probabilité de succès (la probabilité que z_i soit égal à 1).

Nous obtenons alors la valeur suivante pour la fonction de vraisemblance :

$$L(\beta) = \prod_{i=1}^{n}(r(x_{i,1}, \ldots, x_{i,q})^{z_i}(1 - r(x_{i,1}, \ldots, x_{i,q}))^{1-z_i})$$

Or, la fonction de lien étant : $g(x) = ln(\frac{x}{1-x})$, nous avons :

$$\begin{aligned} g(E[Y_i]) &= ln(\frac{E[Y_i]}{1 - E[Y_i]}) \\ &= ln(\frac{1 \times P[Y_i = 1] + 0 \times P[Y_i = 0]}{1 - (1 \times P[Y_i = 1] + 0 \times P[Y_i = 0])}) \\ g(E[Y_i]) &= ln(\frac{P[Y_i = 1]}{1 - P[Y_i = 1]}) \end{aligned}$$

Estimation des paramètres et des intervalles de confiance

Afin d'obtenir une estimation des paramètres de notre modèle, nous partons des résultats obtenus plus haut et nous obtenons :

$$
\begin{aligned}
ln(\frac{P[Y_i=1]}{1-P[Y_i=1]}) &= \beta_0 + \sum_{j=1}^{q} \beta_j x_j \\
\frac{P[Y_i=1]}{1-P[Y_i=1]} &= \exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j) \\
P[Y_i=1] &= (1-P[Y_i=1)\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j) \\
P[Y_i=1] &= \exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j) - P[Y_i=1]\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j) \\
P[Y_i=1](1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)) &= \exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j) \\
P[Y_i=1] &= \frac{\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}
\end{aligned}
$$

Nous avons alors :

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} (\frac{\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)})^{z_i} \times (\frac{1}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)})^{(1-z_i)} \\
ln(L(\beta)) &= \sum_{i=1}^{n} z_i ln(\frac{\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}) + \sum_{i=1}^{n} (1-z_i) ln(\frac{1}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}) \\
ln(L(\beta)) &= ln(\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)) \sum_{i=1}^{n} z_i - n \times ln(1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)) \\
ln(L(\beta)) &= (\beta_0 + \sum_{j=1}^{q} \beta_j x_j) \sum_{i=1}^{n} z_i - n \times ln(1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j))
\end{aligned}
$$

Pour trouver l'estimateur du maximum de vraisemblance d'un paramètre, nous devons trouver les valeurs de β qui maximisent la fonction log-vraisemblance.
Nous devrions donc résoudre le système d'équations suivant afin d'obtenir les estimateurs

pour chaque paramètre :
$$\begin{cases} \frac{\partial ln(L(\beta)}{\partial \beta_1} & = \sum_{i=1}^{n} x_{i,1}(z_i - \frac{\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}) \\ \frac{\partial ln(L(\beta)}{\partial \beta_2} & = \sum_{i=1}^{n} x_{i,2}(z_i - \frac{\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}) \\ \vdots & \\ \frac{\partial ln(L(\beta)}{\partial \beta_q} & = \sum_{i=1}^{n} x_{i,q}(z_i - \frac{\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{q} \beta_j x_j)}) \end{cases}$$

La résolution de ce système étant impossible avec les méthodes analytiques, nous sommes contraints d'utiliser d'autres méthodes pour y parvenir. Nous décidons d'utiliser l'algorithme de **Newton-Raphson** [20] qui est l'un des plus connus.

Celui-ci est mis en place de la manière suivante :

— Tout d'abord, nous choisissons une valeur initiale pour β_0,

— Ensuite, nous posons pour $j,k \in \{0,\dots,q\}$:
$\beta_{j+1} = \beta j - ln(L(\beta_j)(\frac{\partial^2 ln(L(\beta)}{\partial \beta_j \partial \beta_l}))^{-1}$,

— Et enfin, nous arrête l'algorithme dès que $|\beta_{j+1} - \beta_j|$ devient inférieure à un certain seuil choisi par avance.

Cet algorithme nous permet d'obtenir le vecteur $\hat{\beta}$ qui est composé des estimateurs du maximum de vraisemblance des paramètres β_j. Vu que les éléments de $\hat{\beta}$ sont à valeur dans $\mathbb{R}$, nous avons : $\hat{\beta} \sim \mathcal{N}(\mu = \beta, \sigma^2 = (\frac{\partial^2 ln(L(\beta)}{\partial \beta \partial \beta'}))^{-1})$. Cela nous permet de déterminer aisément les intervalles de confiance pour chacun des paramètres :

$$Intervalle \approx \left[(-\hat{\beta} - Q_{1-\alpha})\left(\sqrt{(\frac{\partial^2 ln(L(\beta)}{\partial \beta \partial \beta'}))^{-1})}\right)^{-1}, (-\hat{\beta} + Q_{1-\alpha})\left(\sqrt{(\frac{\partial^2 ln(L(\beta)}{\partial \beta \partial \beta'}))^{-1})}\right)^{-1}\right] \tag{2.2}$$

Avec $Q_{1-\alpha}$ le quantile de niveau $1-\alpha$ pour une variable aléatoire de loi normale centrée réduite. **NB : Normalement, nous avons β au lieu de $\hat{\beta}$ dans la formule, mais vu que nous n'avons pas les valeurs de β, nous utilisons les estimateurs pour avoir une approximation des intervalles de confiance.**

2.2.3 Amélioration des performances de la régression logistique

Une fois toutes ces étapes franchies, nous nous intéressons aux différentes méthodes utilisées afin d'améliorer notre modèle.

Nous devons sélectionner les variables explicatives qui nous permettent d'obtenir le modèle le plus optimal en considérant un critère choisi.

Il existe un nombre fini de modèles possibles vu que le nombre de variables explicatives est fini et pour sélectionner le meilleur d'entre eux, nous pouvons utiliser les méthodes ci-dessous [4] :

— La méthode **forward** : cette méthode consiste à partir d'une régression logistique vide (ne comprenant aucune variable explicative) et à rajouter à chaque étape une nouvelle variable explicative qui améliore le plus les performances du modèle, et ce, jusqu'à ce que l'ajout d'une autre variable aléatoire n'améliore pas le modèle.

— La méthode **backward** : elle consiste à faire le contraire de la méthode forward, nous partons du modèle complet (comprenant toutes les variables explicatives de notre jeu de données) et à retirer à chaque étape la variable ayant le moins d'impact sur la qualité du modèle. La procédure se poursuit jusqu'à ce qu'aucune autre variable explicative ne puisse être retirée sans nuire significativement à la qualité du modèle.

— La méthode **stepwise** : Il s'agit d'un croisement entre les deux méthodes précédentes, elle consiste à partir soit d'un modèle vide (forward stepwise) soit d'un modèle comprenant toutes les variables (backward stepwise) et à chaque étape, ajouter ou supprimer une variable explicative en fonction d'un critère de qualité, puis prendre la décision la plus appropriée. Comme pour les autres méthodes, elle s'arrête lorsque l'ajout ou la suppression d'une variable donnée n'améliore pas la qualité du modèle.

Parmi tous les critères qui sont généralement utilisés, nous en citerons 3 qui permettent de mettre en place les méthodes de sélection des modèles expliquée plus haut et 4 autres qui permettent de vérifier que les prédictions du modèle sont bonnes.

Les critères utilisés pour les méthodes forward, backward et stepwise sont les suivants :

— La **déviance** : elle permet de mesurer la différence entre le modèle ajusté et le modèle de référence. Plus elle est faible et meilleur est l'ajustement du modèle aux données. Pour le modèle de régression logistique, elle est calculée comme suit :

$$DEV(r|Z) = -2\sum_{i=1}^{n}(Z_i ln(r_i(\beta)) + (1-Z_i) ln(1-r_i(\beta)))$$

$$DEV(\beta|Z) = -2\sum_{i=1}^{n}(Z_i(ln(r_i(\beta)) - ln(ln(1-r_i(\beta)))) + ln(1-r_i(\beta)))$$

$$DEV(\beta|Z) = -2\sum_{i=1}^{n}(Z_i(logit(r_i(\beta))) + ln(1-r_i(\beta)))$$

(Car r qui est la probabilité de succès dépend de β)

— Le **critère d'information d'Akaike** ou **AIC** qui se calcule à l'aide de la formule suivante :

$$AIC = -2ln(L(\hat{\beta})) + 2q$$

Où $L(\hat{\beta})$ est la vraisemblance maximisée et q le nombre de variables explicatives du modèle. L'AIC représente donc un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètres possible).

— Le **critère Bayes Information criterion** qui se calcule comme suit :

$$BIC = -2ln(L(\hat{\beta})) + ln(n)q$$

Où n est le nombre de lignes de notre modèle.

Les critères permettant de vérifier que les prédictions des modèles [1] sont bonnes sont les suivants :

— Le **RMSE** ; on ne l'utilise pas pour les variables binaires, mais vu qu'il nous servira pour le traitement des valeurs manquantes, nous préférons le mentionner. Il s'agit d'une métrique qui permet de calculer l'écart entre les valeurs réelles et les prédictions du modèle. La formule pour le calculer est la suivante :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}$$

Plus il est faible et plus le modèle est performant.

— L'**accuracy**, c'est le quotient entre le nombre de prédictions correctes et le nombre total d'observations (ou de valeurs à prédire).

$$Accuracy = \frac{N_{correctes}}{N_{Total}}$$

Où :
$N_{correcte}$ est le nombre de prédictions correctes et N_{Total} le nombre de valeurs à prédire.
Plus il est élevé et plus le modèle est performant.

— La **courbe ROC (*Receiver Operating Characteristic*)** et l'**AUC** : la courbe ROC [12] permet de visualiser la performance d'un modèle de classification en fonction de différents seuils de probabilité, elle correspond à la courbe du taux de vrais positifs en fonction du taux de faux positifs qui est tracée en faisant varier le seuil qui permet de dire si oui ou non un évènement est un succès et l'AUC est l'aire en dessous de cette courbe.
Pour tracer cette courbe, la première étape est de déterminer la **matrice de confusion** du modèle. Cette matrice se définit comme suit :

	$\hat{Z}=1$	$\hat{Z}=0$	Total
$Z=1$	VP	FN	$VP+FN$
$Z=O$	FP	VN	$FP+VN$
Total	$VP+FP$	$FN+VN$	n

Où :

- VP est le nombre de vrais positifs prédits par le modèle,
- FP, le nombre de faux positifs,
- VN, le nombre de vrais négatifs et

1. ces critères sont valables pour les deux autres modèles qui seront abordés

- FN, le nombre de faux négatifs.

Pour obtenir ces éléments, on procède comme suit : Soit s le seuil au-delà duquel on dit que l'évènement est un succès.

— Si $\hat{r} \geq s$ alors $\hat{Z} = 1$ (on rappelle que $\hat{r}$ est la probabilité prédite par le modèle de régression logistique).

— Sinon, $\hat{Z} = 0$.

Une fois cette étape franchie, nous calculons les taux suivants :

— Le taux de vrais positifs (appelé aussi sensibilité) qui est donné par la relation suivante : $\frac{VP}{VP+FN}$; il nous indique la capacité de notre modèle à bien détecter les succès,

— Le taux de faux positifs qui est donné par la relation suivante : $\frac{FP}{FP+VN}$, il nous indique le pourcentage d'échecs prédits correctement par rapport à tous les faux échecs et les vrais échecs prédits. Il est aussi donné par la relation suivante : $1 - specificite$; où la spécificité est la capacité du modèle à bien détecter les échecs.

Ces différents éléments peuvent être utilisés pour déterminer le seuil optimal pour un modèle de régression logistique. En effet, ce seuil peut être déterminé en maximisant la sensibilité et la spécificité du modèle (ou en maximisant la sensibilité et minimisant le taux de faux positifs), ce qui peut être utile si nous voulons spécifier les paramètres de la loi de notre variable réponse par exemple.

2.3 Forêt aléatoire

2.3.1 Théorie

Les forêts aléatoires sont un ensemble d'arbres de décisions, ceux-ci bien qu'étant aisément interprétables, manquent de précision et ne sont pas facilement généralisables, car ils sont la plupart du temps très efficaces sur les données d'entrainement, mais très peu sur les données de test, ce qui est dommage quand on met en place un modèle de machine learning.
C'est dans le but de trouver une solution à ces problèmes que sont utilisées les forêts aléatoires ou *random forest*.

Pour créer les arbres de décision d'un modèle de forêt aléatoire, la plupart du temps ce sont les méthodes ci-dessous qui sont utilisées :

— Le **bagging** qui est la méthode la plus courante et qui consiste à prendre le jeu de données et à entraîner le modèle sur une partie des données et une partie des variables. Chaque arbre de décision va donner un résultat (soit un succès, soit un échec) pour une observation donnée, puis, une décision du *random forest* sera prise en prenant en compte tous ces résultats.

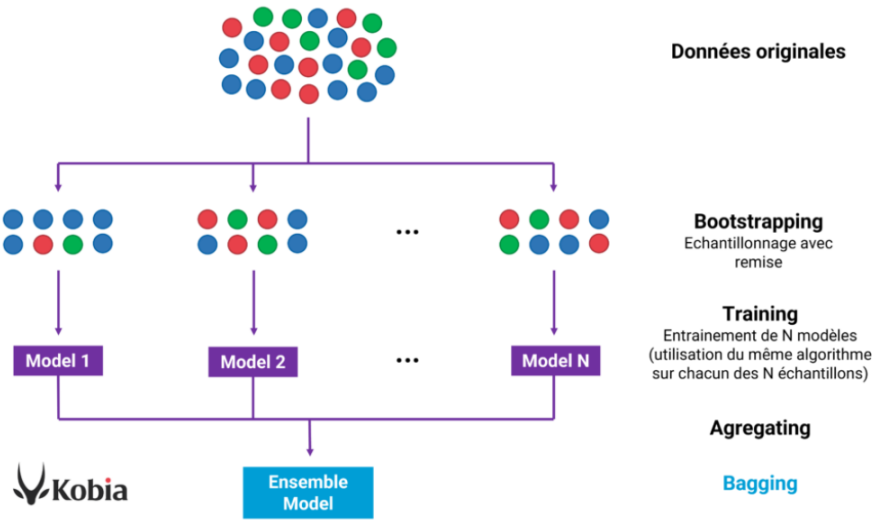


FIGURE 2.1 – Illustration du fonctionnement de la méthode Bagging (Source : [7])

— Le **pasting** qui est très similaire à la méthode dont nous avons parlé précédemment sauf que ce ne sont pas les mêmes jeux de données qui sont tirés pour entraîner les arbres du modèle. Cette méthode est très peu utilisée.

Explication plus approfondie du bagging

Nous allons considérer les mêmes variables explicatives et réponses que nous avions utilisées pour l'explication du modèle de régression logistique.
Nous observons les éléments suivants dans notre jeu de données :

$$\theta = \begin{pmatrix} z_1 & x_{1,1} & \dots & x_{1,m} \\ \vdots & \vdots & & \vdots \\ z_n & x_{n,1} & \dots & x_{n,m} \end{pmatrix}$$

Soit M un modèle (ici un arbre de décision) et θ_b un échantillon *bootstrap* d'observations et M_{θ_b} l'ajustement du modèle M aux données de θ_b.
Afin de réaliser le bagging pour un élément z_u à prédire à partir d'un jeu de données $X_u = (x_{u,1}, \dots, x_{u,m})$, on considère B échantillons bootstrap de θ et Nous procédons comme suit :

— Nous partons d'abord d'un ensemble vide N,

— Puis, on met en place l'algorithme suivant :

- Pour $b \in \{1, \dots, B\}$, nous réalisons le tirage d'un échantillon bootstrap θ_b de θ,
- Nous ajustons ensuite $\hat{M}_{\theta_b}$,
- Puis, on détermine une estimation $\hat{M}_{\theta_b}(X_u)$,
- On pose ensuite $L := L$ et,
- A la fin de l'algorithme, on renvoie le résultat suivant : $\hat{M}_{\theta_b}(X_u) = \frac{1}{B} \sum_{k \in L} L_k$

2.3.2 Méthodes d'amélioration des performances du modèle

Il existe différentes méthodes utilisées pour améliorer des performances des modèles de forêt aléatoire basés sur la recherche d'hyperparamètres optimaux.

Ceux que nous utiliserons dans notre étude sont :

— La méthode de recherche des hyperparamètres par quadrillage ou encore ***Grid Search*** : cette méthode consiste à prendre une liste de possibilités pour chacun des hyperparamètres du modèle puis de réaliser une combinaison de ces listes et enfin pour chaque combinaison entraîner le modèle puis calculer son score.
Nous conservons à la fin les hyperparamètres ayant donné le modèle avec le meilleur score. Cet outil est assez intéressant, car il permet de réaliser une sélection assez fine des hyperparamètres (cela peut nous permettre de mettre en place des représentations graphiques des scores obtenus et donc de mieux visualiser nos résultats) mais il représente un énorme désavantage du fait de son caractère très chronophage.

— La méthode de recherche aléatoire des hyperparamètres ou *Randomized Search* : elle permet comme la méthode *grid search* de trouver les meilleurs hyperparamètres pour notre modèle, l'outil utilisé pour ce faire fonctionne de manière similaire au *grid search* mais recherche de manière aléatoire des combinaisons d'hyperparamètres,

— La méthode d'optimisation Bayésienne aussi appelée *Optuna Search* qui permet de réaliser les mêmes choses de manière plus optimale, comme pour le randomized search, l'outil choisit de manière aléatoire des combinaisons d'hyperparamètres des différentes listes au départ puis elle réalise les autres essais en prenant en compte les résultats obtenus précédemment.

2.3.3 Importance des variables

Lorsque l'objectif de l'étude ne se limite pas à évincer des variables du modèle ou à retenir uniquement les variables explicatives essentielles pour garantir une prédiction précise, nous pouvons être enclins à réaliser une hiérarchisation complète de toutes les variables, basée sur leur influence relative par rapport à la variable réponse. Ainsi, ce critère d'importance offre un classement des variables, allant de la plus significative à la moins significative.

Pour avoir une idée de l'importance de nos variables explicatives, nous utilisons l'accroissement moyen de l'erreur d'un arbre de la forêt aléatoire lorsqu'on réalise une permutation dans les échantillons d'observations θ_b de manière aléatoire.

Nous trouverons ci-dessous les étapes détaillées du calcul du coefficient d'importance pour une variable explicative X_j avec $j \in \{1,\dots,m\}$:

— Tout d'abord, nous considérons l'échantillon bootstrap b de θ et θ_b l'échantillon d'observations qui lui est associé. Nous allons calculer $error\theta_b$ qui est l'erreur sur l'échantillon θ_b commise par l'arbre construit sur b.

— Ensuite, nous permutons de manière aléatoire les valeurs de la variable explicative X_j toujours dans l'échantillon θ_b et calculons l'erreur $error\hat{\theta}_b^j$. cette étape est réalisée pour tous les échantillons bootstrap.

— Et enfin, nous déterminons le coefficient d'importance de X_j, $Imp(X_j)$ qui représente la différence entre l'erreur moyenne de l'arbre cité plus tôt sur l'échantillon permuté $\hat{\theta}_b^j$ et celle de ce même arbre sur l'échantillon initial θ_b.

  Ce coefficient est donné par la relation suivante : $Imp(X_j) = \frac{1}{m}\sum_{l=1}^{m}(error\hat{\theta}_b^j - error\theta_b)$

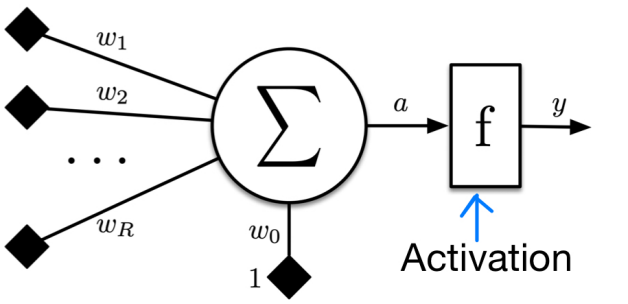
L'importance la variable évolue dans le même sens que l'erreur engendrée par la permutation aléatoire de celle-ci. Si les permutations de la variable ont un impact considérable sur l'erreur, alors ladite variable est importante pour le modèle, dans le cas contraire cela témoigne de son caractère non significatif [22].

2.4 Réseaux de neurones

2.4.1 Théorie

Inspirés de l'architecture des réseaux neuronaux du cerveau, les réseaux de neurones artificiels sont un ensemble de fonctions statistiques permettant d'effectuer différentes tâches en intelligence artificielle.

L'unité de ces réseaux est donnée par le modèle du *"threshold logic unit" (TLU)* représentant un neurone. Dans sa forme la plus simple, un neurone est un discriminant linéaire avec une fonction d'activation dont les entrées et sorties sont des nombres. Un réseau de neurones est formé de plusieurs neurones connectés entre eux.
On distingue plusieurs types de réseaux de neurones, notamment, ceux à une couche dans lesquels tous les neurones sont connectés à toutes les entrées, appelés perceptrons. Et les perceptrons multicouches dans lesquels les sorties des neurones d'une couche i sont les entrées de la couche i+1, chaque couche fait une projection dans un nouvel espace et l'information se propage des entrées vers les sorties.



$$y = \mathrm{f}\left(\sum_i w_i x_i + w_0\right) = \mathrm{f}(\mathbf{w}^\top \mathbf{x} + w_0)$$

- Fonction linéaire : $\mathrm{f}_{lin}(a) = a$
- Fonction sigmoïde : $\mathrm{f}_{sig}(a) = \frac{1}{1+\exp(-a)}$

FIGURE 2.2 – Précision sur le fonctionnement des réseaux de neurones (Source : [14])

2.4.2 Rétropropagation

Dans un perceptron multicouche, la ou les premières couches sont appelées couches cachées et la dernière est la couche de sortie. L'entraînement de ce type de modèle consiste

au passage de données d'une couche à une autre en utilisant une combinaison de la méthode de la descente du gradient et de la dérivation automatique en mode inverse (en anglais *reverse-mode automatic differenciation*) qui calcule le gradient de l'erreur d'un réseau neuronal en considérant tous les paramètres du modèle (hyper-paramètres).

Dans la rétro propagation, premièrement, on entraîne le réseau sur plusieurs sous-ensembles des données. Ensuite, chaque sous-ensemble est passé à la couche d'entrée et les résultats sont passés aux couches suivantes jusqu'à la couche de sortie. Dans cette passe avant, toutes les prédictions intermédiaires sont conservées. Par la suite, l'algorithme mesure l'erreur du modèle en comparant la sortie désirée à la sortie prédite, il évalue à quel point les biais de chaque sortie précédente et des connexions à la couche de sortie ont contribué à cette erreur. La prochaine étape consiste en une passe en arrière, dans laquelle l'algorithme permet de déterminer dans quelle mesure ces contributions d'erreurs proviennent de chaque connexion aux couches cachées. Finalement, l'algorithme effectue les étapes de la descente du gradient pour ajuster les poids dans le réseau en utilisant les erreurs précédemment calculées [15].

2.4.3 Importance des variables ou caractéristiques

Dans le contexte des réseaux de neurones, établir l'importance des variables est crucial pour saisir quelles entrées contribuent le plus à la prédiction du modèle. Le *Gradient Feature Importance* se base sur la notion que si une variable particulière est significative, son impact sur la sortie (la perte) du réseau sera plus prononcé.

Mathématiquement, cela implique le calcul des dérivées partielles des sorties par rapport aux variables d'entrée. En utilisant la règle de la chaîne, la dérivée partielle de la sortie par rapport à une variable donnée peut être décomposée en produit de dérivées partielles locales, chacune correspondant à une couche du réseau. En effectuant ce calcul pour toutes les variables, nous obtenons une vue globale de leur influence.

Plus précisément, si la variation d'une variable conduit à une variation significative de la perte, c'est un indicateur fort de son importance. Inversement, si une modification n'a qu'un effet marginal, voire négatif sur la perte (c'est-à-dire si l'erreur diminue), cela signale que la variable a moins d'impact.

Chapitre 3

Données et traitement des valeurs manquantes

3.1 Présentation des données

Pour mener à bien notre étude, nous disposions de deux sources de données de la prévoyance individuelle AXA France :

— Les **tables brutes CRM** : CRM est le point de départ de toute demande d'indemnisation en arrêt de travail chez AXA France, c'est sur cette plateforme que les gestionnaires de sinistre vont renseigner les différentes informations relatives à un nouveau dossier quelle que soit la décision ce qui fait que nous y retrouvons les informations sur tous les dossiers, qu'ils soient après traitement acceptés ou refusés. Ces données sont brutes et certaines informations peu fiables, car plusieurs champs de la plateforme CRM sont remplis manuellement par les gestionnaires, elles nécessitent donc un prétraitement avant d'être utilisées.
Dans le cadre de ce mémoire, les tables brutes qui nous intéressent sont :

- La table brute **dossier** qui contient des informations sur chaque nouveau dossier, elle a la vue une ligne par dossier et a des variables telles que le type de dossier, le statut, ou encore le motif du statut du dossier,
- La table brute **contrat** qui elle contient les informations relatives aux contrats, elle a la vue une ligne par contrat après traitement et contient les variables telles que la date d'effet du contrat, ou encore la situation du contrat,
- La table brute **sinistre** qui contient les informations relatives à chaque synthèse de sinistre (regroupement d'ouvertures et de prolongations d'arrêt de travail ayant le même évènement générateur), elle a la vue une ligne par synthèse sinistre après traitement et contient des variables telles que la date de création, numéro de synthèse sinistre ou encore la date de survenance du sinistre,
- La table brute **commentaire** qui contient les commentaires laissés par les gestionnaires dès l'arrivée d'un nouveau dossier sur la plateforme CRM, elle a la vue une ligne par dossier et des variables telles que le numéro du dossier ou encore les commentaires pour chaque dossier,
- Et enfin la table brute **compte** qui contient des informations sur les assurés, elle a une vue une ligne par personne après traitement et des variables telles que les coordonnées de l'assuré (nom, prénom, adresse postale…), ou encore sa profession.

— Nous avons aussi les bases de données de **VANDA** qui elles sont très propres et fiables, car les champs de VANDA sont pour la plupart automatisés. Bien qu'étant très propres et directement exploitables, ces données contiennent les informations relatives au sinistre (date de survenance, cause…) seulement pour les dossiers qui ont été acceptés, ce qui fait qu'elles ne suffisent pas pour la réalisation de notre analyse.
Il en existe plusieurs mais dans le cadre de ce mémoire, nous nous sommes intéressés aux bases VANDA suivantes :

- La base **contrat** qui contient les informations relatives aux contrats,

- La base **garantie** qui contient toutes les informations relatives aux différentes garanties telles que les nombres de jours des franchises accident, maladie et hospitalisation,
- La base **personne** qui contient les informations sur l'assuré telles que sa date de naissance ou encore une variable qui dit s'il est fumeur ou non,
- La base **surprime** dont les principales informations sont le fait que l'assuré paie une surprime pour son contrat ou ne bénéficie pas d'une réduction,
- La base **sinistre** qui contient toutes les informations relatives aux sinistres telles que la date de début du sinistre et sa date de fin...
- Et enfin, une deuxième base **garantie** qui elle contient les informations relatives aux garanties que nous n'avons pas dans la première telle que le montant des indemnités journalières pour chaque garantie.

Dans la suite, nous parlerons de manière plus détaillée de ces données, des différentes étapes du traitement et de leur utilité.

3.2 Utilisation des méthodes de Text Mining pour la création de nouvelles variables

3.2.1 Présentation du text mining

Le *text mining* est un ensemble de méthodes informatiques permettant d'extraire des informations comprises dans des ensembles de données textuelles non-structurées (brutes) ou semi-structurées.

Ces méthodes permettent aux entreprises qui les utilisent de pouvoir exploiter des données issues de sources diverses, cela peut être par exemple les données issues de la comptabilité ou du CRM (le cas auquel nous sommes confrontés dans notre étude) qui comme elles sont textuelles et renseignées différemment selon la personne manquent beaucoup de structure.

En général, l'utilisation de ces méthodes nécessite le suivi de différentes étapes [1] qui sont les suivantes :

— Le prétraitement des textes : cette étape consiste à rendre un peu plus homogène nos textes. Cela se fait selon le processus suivant :
  - La suppression des ponctuations et des *stopwords* (mots courants) tels que du, des, les,...
  - La mise en majuscule de tous les mots de nos textes,

1. le nombre d'étapes peut varier en fonction des objectifs, ici nous citerons les étapes qui nous ont servi pour l'étude.

  - La création d'un "dictionnaire" contenant les mots ayant la même racine (les abréviations et les mots complets). Nous pourrions par exemple avoir dans notre cas le dictionnaire suivant :
    dico = {INCAP : INCAPACITE, HOSPI : HOSPITALISATION}

— L'extraction des informations : pour cette étape, plusieurs possibilités s'offrent à nous selon ce que nous souhaitons faire. Dans notre cas par exemple, vu que nous voulions utiliser les méthodes de text mining pour créer de nouvelles variables, nous avons créé des fonctions permettant de récupérer l'information ciblée dans le texte homogénéisé.

3.2.2 Jointure des tables CRM importantes

Nous utiliserons 5 bases de données brutes importantes que nous allons fusionner dans un premier temps comme mentionné précédemment.
Ce sont :

— La base **dossier** qui contient des variables telles que : Type, Statut, Motif du statut, Date de création, Date de clôture, Origine, Objet, Commentaires éventuels, Événement associé, Importance du client, TOP Analyse Complémentaire...

— La base **contrat** qui contient des variables telles que : Code produit, Date d'effet, Situation, Fractionnement, Numéro de contrat, Cotisation TTC, Cumul des primes versées, Date de résiliation...

— La base **sinistre** qui contient des variables telles que : Date de création, Numéro de sinistre CRM (pas le même que celui de VANDA), Date de survenance, TOP Multi contrat, Événement générateur, Typologie...

— La base **compte** qui contient des variables telles que : Coordonnées (nom, prénom, genre, e-mail, téléphone, adresse du foyer), rôle dans le foyer, Type de client, Code profession (CSP), Profession exacte, Situation familiale, Type de contrat de travail... et

— La base **commentaire** qui contient par exemple le contenu du commentaire dossier et la dernière date de modification.

Ces tables contiennent toutes les informations qui vont nous permettre de qualifier de manière précise une nouvelle demande d'indemnisation en arrêt de travail, la base sinistre par exemple nous apportera des informations beaucoup plus précises sur l'évènement ayant provoqué l'arrêt, à elles toutes, elles représentent donc un flux de données très important. Pour construire notre base de donnée, nous utilisons les vues les plus complètes de ces données, elles contiennent donc toutes les informations disponibles sur CRM depuis sa mise à disposition chez AXA France et nous avons gardé pour chaque table, seulement les variables qui étaient non triviales (les variables non-nulles, possédant plus d'une modalité et avec un taux de complétude assez important) et pour pouvoir les joindre, nous avons dû procéder étape par étape, car elles ne contenaient pas toutes les mêmes variables **clés**.

— Nous avons dans un premier temps joint les tables dossier et contrat en utilisant les clés "N_contrat_AXA" et "Id" qui contiennent la même information provenant respectivement des tables citées,

— Puis, la table dossier à la table compte en utilisant comme clé la variable "AccountId",

— Nous avons ensuite joint les tables dossier et sinistre en utilisant les variables "synthèse_du_sinistre" et "SinistreId" qui, elles aussi, contiennent la même information (le numéro de la synthèse sinistre) et,

— Enfin, nous avons joint la table dossier à la table commentaire en utilisant "CaseId" de la table dossier et "ParentId" de la table commentaire.

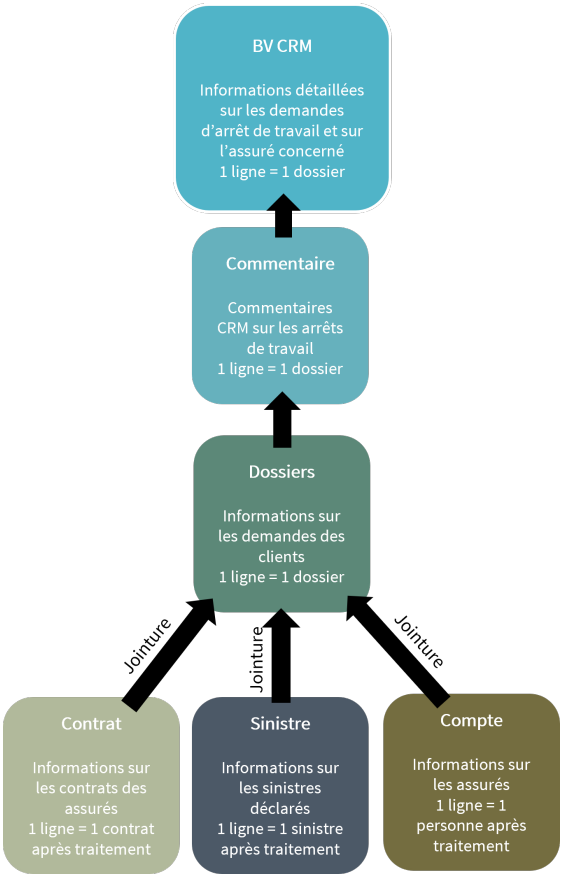


FIGURE 3.1 – Schématisation des jointures réalisées pour créer la BV CRM

Afin d'alléger le volume des données et d'en faciliter les analyses, nous avons restreint notre périmètre d'étude.

— Dans un premier temps, nous avons appliqué un filtre temporel : nous ne considérons que les dossiers ouverts après le 18 mai 2021, car c'est à cette date qu'a été créée la synthèse sinistre de la plateforme CRM,

— Puis, nous avons filtré la base pour garder uniquement les dossiers d'ouverture d'arrêt de travail : nous ne nous intéressons qu'aux commentaires de l'ouverture, car le but de notre étude est de pouvoir prédire l'acceptation ou le refus d'une demande d'indemnisation pour une **ouverture d'arrêt de travail**, et,

— Enfin, nous avons appliqué un filtre sur les produits auxquels nous nous intéressions : Produit 1, Produit 2, Produit 3 et Produit 4.

Nous avons aussi créé une nouvelle variable que nous appellerons "OUTCOME" qui dit si oui ou non le dossier est accepté. La création de cette variable s'est faite suivant les différentes étapes que nous retrouverons ci-dessous :

— Nous observons de plus près les modalités de la variable "statut" de la base sinistre. Le statut ne prend que deux modalités qui sont "Refus" et "Clôturé" cependant, une première étude ayant démontré que tous les dossiers avec le statut "Refus" sont bien refusés, mais pas tous ceux qui ont le statut "Clôturé" sont acceptés, nous décidons à la suite de cette première observation d'observer de plus près la variable "motif_statut" afin de mieux comprendre le statut "Clôturé".

— Cette variable nous permet d'avoir plus de détails sur la raison expliquant le statut du sinistre lorsque le statut est "Clôturé", elle a des modalités telles que :

  • Franchise : qui dit que le dossier est clôturé parce que la franchise est atteinte, ce qui laisse entendre qu'il a été refusé,

  • Pas de garantie souscrite : qui comme son nom l'indique dit que le dossier est clôturé, car l'assuré n'a pas de garanties IJ rattachées à son contrat, ce qui laisse entendre qu'il a été refusé,

  • Clôturé règlement : qui dit que le dossier est clôturé, car nous avons décidé de payer pour le sinistre concerné, ce qui laisse entendre qu'il a été accepté et qu'il y aura indemnisation, ...

— Après avoir observé tous les motifs de statuts de sinistres, nous décidons de qualifier comme dossiers acceptés les dossiers ayant pour motifs : "clôturé traité", "clôturé règlement", "Réclamation traitée", "clôturée réclamation traitée" et "clôturé accord de prise en charge". Les dossiers ayant des motifs ne figurant pas dans cette liste sont qualifiés comme étant refusés.

Nous obtenons, après ces différents traitements, une nouvelle base que nous appellerons CRM, qui a environ 60 000 lignes (une ligne correspondant à un commentaire de CRM pour un dossier donné) ce qui représente 25 000 dossiers au 30 avril 2023 et nous observons sur nos dossiers la constituant, 18 000 acceptés et 7 000 refusés. Notre base contient les variables que vous trouverez dans le tableau ci-dessous :

Nom de la variable	Descriptif	Table d'origine
NUMERO_CONTRAT	Numéro du contrat sinistré	Compte
NUMERO_DOSSIER	Numéro du dossier de CRM	Dossier
NUMERO_PORTEFEUILLE	Numéro du portefeuille auquel appartient le contrat	Compte
STATUT	Statut du contrat qui peut être "Refus" ou "Clôturé"	Dossier
MOTIF_STATUT	Détails sur les raisons derrière le "STATUT" du contrat, peut-être "Clôturée traitée", "Clôturé Règlement",...	Dossier
DATE_CREATION_DOSSIER	Date de création du dossier sur CRM	Dossier
RESEAU	Réseau de distribution du contrat auquel est rattaché le dossier	Dossier
OBJET_DEMANDE	Objet de la demande qui peut être par exemple "Ouverture incapacité" ou encore "Prolongation arrêt de travail"	Dossier
IMPORTANCE_CLIENT	Indique le niveau de fidélité de l'assuré	Dossier
COMMENTAIRE	Commentaires CRM rédigés par les gestionnaires et associés à un dossier donné	Commentaire
CODE_UV	Code produit	Contrat
DATE_EFFET_CONTRAT	Date d'effet du contrat	Contrat
NUMERO_CONTRAT	Numéro du contrat	Contrat
NOM_PRODUIT	Nom du produit concerné qui peut être "Produit 1", "Produit 2", "Produit 3" ou "Produit 4"	Contrat
SITUATION_CONTRAT	État du contrat, peut-être "en cours", "résilié",...	Contrat
DATE_RESILIATION_CONTRAT	Date de résiliation du contrat si la "SITUATION_CONTRAT" indique qu'il est résilié	Contrat
DATE_CREATION_SINISTRE	Date de création de la synthèse sinistre associée au dossier sur CRM	Sinistre
NUMERO_SINISTRE	Numéro de la synthèse sinistre à laquelle appartient le dossier	Sinistre
DATE_SURVENANCE_SINISTRE	Date de survenance de l'évènement ayant généré le sinistre	Sinistre

TABLE 3.1 – Récapitulatif des variables de notre base de données (1/2)

Nom de la variable	Descriptif	Table d'origine
CAUSE_SINISTRE	Cause du sinistre, peut prendre les modalités "accident", "maladie"	Sinistre
ROLE_FOYER	Place de l'individu dans le contrat, peut-être "assuré","co_assuré",...	Compte
CODE_CSP	Code de la profession exercée par l'assurée	Compte
PROFESSION	Profession de l'assuré	Compte
SITUATION FAMILIALE	Situation familiale de l'assuré, peut-être "célibataire", "marié",...	Compte
PRESENCE_ENFANT	Présence d'enfants au sein du foyer, peut-être "true" ou "false"	Compte
SEXE	Sexe de l'assuré	Compte
TYPE FLUX	Le moyen par lequel les différents éléments du dossier parviennent à l'assureur, courrier, espace client ...	Dossier

TABLE 3.2 – Récapitulatif des variables de notre base de données (2/2)

3.2.3 Création de nouvelles variables à partir des commentaires de la base CRM

Les commentaires de CRM sont très importants pour l'étude, car ils contiennent des informations sur le contrat et les garanties souscrites, par exemple au moment de la survenance même du sinistre, mais aussi des informations très importantes sur le sinistre comme les dates de début et de fin d'arrêt de travail, ou encore les dates de début et de fin d'hospitalisation ce qui fait que l'information contenue dans ces commentaires est récente par rapport à la date de survenance du sinistre.
Nous avions au début décidé de garder uniquement le premier commentaire de l'ouverture pour chaque dossier (25 000 au total, au 20/01/2023). Mais après traitement des données, nous nous sommes rendu compte que nous arrivions à extraire certaines informations pour moins de 30% des dossiers, ce qui représente environ 7 500 dossiers. Nous avons donc décidé de fusionner tous les commentaires de l'ouverture pour chaque dossier (afin de garder le maximum d'informations possibles dans la suite de nos analyses) puis, de les transformer en liste, ce qui facilitera l'application des méthodes de texte mining sur ceux-ci et donc la création de nouvelles variables.

i) Extraction des dates de début et de fin de l'arrêt de travail des commentaires CRM

Les dates de début d'AI

Les commentaires étant rédigés manuellement par différents gestionnaires, lors des premières manipulations des données, nous avons remarqué un manque de structure, ce qui

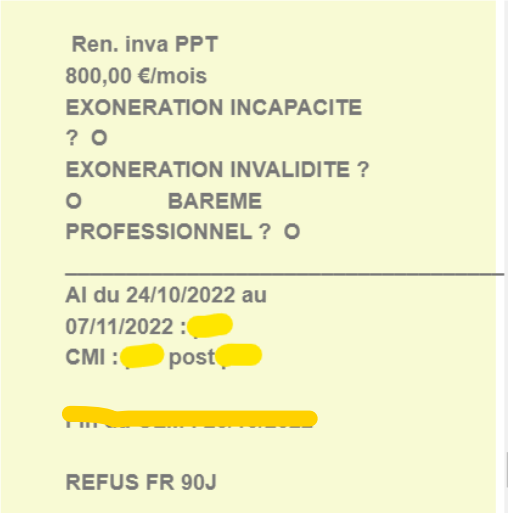


FIGURE 3.2 – Exemple de commentaire sur CRM avant traitement

rendait difficile l'extraction immédiate des dates sans les restructurer un minimum au préalable.
Nous avons donc dû suivre les étapes suivantes pour traiter ces dates :

— Dans un premier temps, nous avons essayé d'homogénéiser la structure des différents commentaires dont nous disposions en remplaçant toutes les lettres en minuscule par des majuscules, les apostrophes par des espaces, et en supprimant toutes les ponctuations (on en a aussi profité pour corriger quelques fautes de frappes que nous avions remarquées, comme des dates qui étaient collées à des mots tels que "au", "du", "et" ...).

— Puis, nous avons créé un dictionnaire avec des mots clés « synonymes » car ils étaient écrits de différentes façons par chaque gestionnaire. Certains écrivaient par exemple « incap » ou « inca » pour parler d'incapacité ou encore « acc » pour accident et nous avons remplacé ces mots par un seul et même mot dans chacun des commentaires.

— Enfin, nous avons créé une liste de *stopwords* (mots vides de sens) tels que « le », « du », « au »... et nous les avons tous supprimés des commentaires fusionnés.

Nous avons après ces premiers traitements mis en place certaines fonctions pour faciliter l'extraction des dates et avons procédé en suivant les étapes ci-dessous.

1. Nous avons créé une fonction permettant de dire si oui ou non une chaîne de caractères peut être transformée en date. Au début, cette fonction ne prenait en compte que les formats de dates avec 4, 5, 6, 7 et 8 chiffres, mais nous nous sommes rendu compte après avoir observé un peu plus en profondeur les commentaires que ce n'étaient pas les seuls formats de dates que nous avions, nous avons donc décidé de créer une autre fonction pour dire si oui ou non une chaîne de caractère peut être

transformée en nombre entier à 1 ou deux chiffres pour prendre en compte les cas où la seule information contenue dans les dates était le jour.

2. Nous avons ensuite créé plusieurs listes de mots « clés » que nous retrouvions très souvent dans le commentaire juste avant de trouver la date de début contenant des mots tels que « AI », « AT », « AISPE », « ITT », « SIN » ou encore « IJ »,
3. Puis, nous avons créé plusieurs fonctions d'extraction des dates fonctionnant de telle sorte que nous parcourons le commentaire jusqu'à l'avant-dernier mot et si nous trouvons un mot des différentes listes et que juste après, nous avons une date, la date de début est cette dernière. Nous avions au préalable classé ces listes par ordre d'importance en quelque sorte, nous avons les mots clés les plus "sûrs" (ceux pour lesquels nous sommes pratiquement sûrs que si nous avons une date juste après, il s'agit forcément de la date de début de l'arrêt) dans la première liste et ainsi de suite. Nous en avons créé finalement 5 qui nous ont permis d'extraire les dates pour plus de 70% des dossiers.
4. Et enfin, nous avons ciblé les 30% de dossiers restants (7 500 dossiers) pour lesquels nous n'avions toujours pas pu extraire les dates de début de l'AI, et nous avons remarqué que dans la plupart des commentaires de ces dossiers, nous n'avions tout simplement pas les dates. Nous avons donc décidé d'ajouter le premier commentaire de la première prolongation pour les dossiers qui en ont au commentaire fusionné de l'ouverture, puis, nous y avons appliqué les fonctions d'extraction de dates créées. Nous obtenons 80% de dossiers (soit 20 000 dossiers) pour lesquels nous arrivons à extraire la date.

En ce qui concerne les autres dossiers, une étude nous a permis de nous rendre compte que la plupart ne contenaient pas de dates et les autres aucun mot clé, nous avons donc décidé de nous arrêter là et de ne pas aller plus loin pour l'extraction des dates de début d'arrêt.

Les dates de fin d'AI

Nous partons de nos 20 000 dossiers pour lesquels nous avons pu extraire la date de début de l'AI pour l'extraction de la date de fin.

1. Nous réalisons dans un premier temps une double boucle for qui récupère dans le commentaire (transformé en liste) le premier élément venant juste après la date de début s'il s'agit bien d'une date, puis on récupère les 2 éléments qui viennent juste après la date de début de l'AI dans le cas où après transformation du commentaire en liste, on se retrouve avec une date séparée. On récupère la date pour 70% de dossiers (soit 14 000 dossiers).
2. Et nous nous rendons compte après observation des commentaires restants que les gestionnaires pouvaient dans certains cas renseigner la date de fin de l'AI d'une autre manière. En effet, certains écrivaient « AI jusqu'au » ou encore « reprise le » et/ou « AT initial » ou « prolong » entre la date de début de l'AI et sa date de fin, raison pour laquelle pour finir, nous n'avons pas supprimé le mot « jusqu » de

nos commentaires puisqu'en le supprimant, on risquait de prendre la date de fin de l'AI à la place de la date de début. Nous prenons donc en compte les cas où nous avions des fautes de frappe dans l'écriture de la date (exemple : au22/12/2019 ou bien 03/02/19+prol), pour ce faire, nous avons écrit un code que nous retrouverons en annexe qui est tel que lorsqu'on retrouve ce cas de figure dans le commentaire, nous prenons seulement une partie de l'élément qui vient après la date de début de l'AI (tout sauf les 2 premiers éléments ou tout sauf les 4 derniers éléments en fonction de l'erreur). Nous passons de 70% (14 000 dossiers) de dossiers pour lesquels nous avons récupéré la date de fin à 85% (17 000 dossiers).

Après ce traitement, nous nous sommes intéressés aux différentes raisons pour lesquelles nous n'arrivions pas à extraire toutes les dates de fin pour tous les dossiers et il s'agissait des mêmes que les raisons pour lesquelles nous n'arrivions pas non plus à extraire les dates de début pour tous les dossiers.

ii) Amélioration du format des dates de début et de fin de l'AI et correction des incohérences

Nous décidons pour mener à bien cette étape de récupérer dans la base CRM de départ la date de survenance du sinistre, ainsi que les dates de création et d'ouverture de la synthèse sinistre pour pouvoir compléter les dates pour lesquelles il y a des informations manquantes telles que le mois ou l'année.
Nous procédons comme suit :

1. Après extraction des dates dans le commentaire, nous décidons d'ajouter la date de survenance du sinistre pour pouvoir compléter les dates (rajouter l'année et/ou le mois) qui n'étaient pas au format jj/mm/aaaa. N'ayant pas les dates de survenance pour tous les dossiers (absence pour un peu plus de 2000 dossiers (surtout des dossiers refusés) sur les 20 000 de départ), nous décidons d'ajouter aussi la date de création de la synthèse sinistre ainsi que la date d'ouverture du dossier. Nous créons de nouvelles variables contenant distinctement le jour, le mois et l'année de la date de survenance et de la date de création de la synthèse et on complète dans un premier temps la date de début de l'AI en réalisant les scénarios suivants :
   - Dans le cas où il nous manque seulement l'année (format jj/mm ou j/mm ou jj/m) dans la date de début, nous comparons le mois de début de l'AI au mois de la survenance de l'AI. Si nous nous rendons compte que le mois de la date de début de l'AI est supérieur ou égal au mois de la date de survenance du sinistre, l'année dans la date de début de l'AI est la même que l'année de la date de survenance. Si, au contraire, le mois de la date de début de l'AI vient avant celui de la date de survenance, l'année de début de l'AI est l'année qui vient après celle de la date de survenance.
   - On suit le même raisonnement pour les dates de début de l'AI pour lesquelles il nous manque aussi le mois (format jj ou j) en se basant cette fois sur le jour dans la date de survenance de l'AI.

- Pour les dates au format jj/mm/aa, j/mm/aa, jj/m/aa, j/mm/aaaa et enfin jj/m/aaaa, nous rajoutons « 20 » lorsque nous avons seulement les deux derniers chiffres de l'année et/ou « 0 » lorsqu'il nous manque des 0 (exemple : 7 pour 07) dans le jour ou dans le mois de la date de fin de l'AI.

2. Après avoir amélioré le format des dates de début de l'AI, nous passons aux dates de fin. Nous procédons de la même manière que pour les dates de début d'AI, mais la comparaison se fait entre les dates de début et de fin d'AI :
   — Dans le cas où nous avons la date de fin de l'AI au format j/mm ou jj/m ou jj/mm, nous prenons l'année de la date de début de l'AI si le mois est postérieur à celui de la date de début de l'AI, dans le cas contraire, nous prenons l'année suivante.
   — Dans le cas où nous avons la date au format jj/mm/aa ou j/mm/aa ou jj/m/aa ou j/mm/aaaa ou jj/m/aaaa, nous rajoutons des « 0 » et/ou « 20 » pour obtenir le format jj/mm/aaaa.
3. Après avoir amélioré le format des dates de début et de fin d'AI, nous avons corrigé les dates incohérentes. Nous nous sommes rendu compte que lorsque le début de l'AI a lieu en fin d'année et la fin de l'AI en début de l'année suivante, pour certains dossiers, les gestionnaires se trompaient sur l'année, ils gardaient la même année pour le début et la fin de l'AI. Nous avons donc écrit un algorithme pour apporter des modifications sur ces dates. Nous avons changé ensuite le type des variables date de début et de fin d'AI (de chaîne de caractère à date) et en comparant les dates de début aux dates de fin, nous nous sommes rendu compte que nous avions des durées négatives pour une vingtaine de dossiers, nous avons donc décidé de corriger ces erreurs manuellement.

iii) Extraction des dates de début et de fin d'hospitalisation et création de la variable top hospitalisation

Les dates de début et de fin d'hospitalisation

Nous procédons de la même manière que pour les dates de début et de fin de l'AI sauf que nous avons beaucoup moins de mots clés à repérer, nous nous intéressons uniquement à **"Hospi"**, **"CRH"** et **"BH"**. Nous arrivons à extraire ces dates pour seulement 1 500 dossiers au total (soit, 06% de dossiers sur nos 25 000 de départ).
Nous améliorons ensuite le format des dates extraites en utilisant la même méthode que pour les dates de l'AI, nous apportons manuellement les corrections pour les dossiers pour lesquels la durée est négative, puis enfin, nous changeons le type des variables date de début et de fin de l'hospitalisation.

Le top hospitalisation

Pour créer cette variable, nous procédons comme suit :

1. Nous créons une variable qui n'a que la modalité 'NON' dans un premier temps,

2. Ensuite, nous créons une boucle for qui change cette modalité en 'OUI' si nous avons pu avoir les deux dates d'hospitalisation et,
3. Nous en créons une autre qui change le 'NON' (qui correspond au cas où on n'a ni dates ni mots clés dans le commentaire CRM) en 'OUI sans dates' si nous avons au plus l'une des deux dates ou la présence de l'un des mots clés CRH ou BH dans le commentaire.

Nous avons de l'hospitalisation pour 3 000 dossiers.

iv) Création des variables liées aux différentes garanties

Le but de cette partie était de pouvoir recueillir toutes les informations relatives aux différentes garanties telles que les montants d'IJ garantis, les différentes franchises et les durées maximales d'indemnisation) que nous retrouvons dans le commentaire.
Nous disposons de 7 différentes garanties dans notre base de données. Ce sont :

1. La garantie toute cause,
2. La garantie IJ RO 365 jours,
3. La garantie IJ RO 90 jours,
4. La garantie IJ RFP (Remboursement Frais Professionnels),
5. La garantie complémentaire à la garantie IJ RO 365 jours,
6. La garantie complémentaire à la garantie IJ RO 90 jours et,
7. La garantie IJ en cas d'accident.

Nous utilisons pour ces variables les mêmes principes que pour les autres, sauf que nous les mettons en pratique avec les fonctions UDF de Python qui nous permettent d'appliquer les fonctions directement aux bases de données, sans les faire s'enregistrer individuellement.
Nous arrivons à extraire et donc à créer les variables suivantes pour chacune des garanties citées plus haut :

— Le montant d'IJ,
— La franchise accident,
— La franchise hospitalisation,
— La franchise maladie,
— Et la durée maximale d'indemnisation.

Nous créons aussi une variable **"Exonération"** qui dit si oui ou non l'assuré est exonéré du paiement de ses primes pendant son arrêt de travail.

Après avoir effectué ces différents travaux, nous nous sommes demandés si les différentes méthodes utilisées donnaient des résultats cohérents, pour se rassurer de la cohérence de nos résultats, nous avons donc décidé d'effectuer des comparaisons entre les informations contenues dans les variables de VANDA et celles que nous obtenons avec nos méthodes à partir des données de CRM.

3.3 Vérification de la conformité des méthodes utilisées

3.3.1 Ajout des données VANDA

Nous décidons d'ajouter les données des bases contrat, garantie, personne et surprime de VANDA à toute la base CRM et les données des bases sinistre et garantie stock seulement aux dossiers acceptés de notre base CRM (les 18 000 dossiers acceptés de départ).

En effet, la base sinistre de VANDA ne contient que les dossiers acceptés et la jointure avec la deuxième base garantie ne peut se faire qu'à partir de la variable "Numero_sinistre" qui est seulement dans ces deux bases et n'est renseignée que pour les dossiers acceptés encore une fois.

i) Enrichissement de la base CRM avec la base garantie de VANDA

Nous utilisons une librairie de la prévoyance individuelle AXA France pour pouvoir importer les bases de données VANDA. Elle nous permet d'avoir la vue de ces bases pour une date choisie (dans notre cas, nous avons décidé de prendre les vues du 20/03/2023, car nous pensons qu'il faut tout au plus 3 mois pour ajouter un nouveau dossier accepté dans VANDA. La BV Garantie a une vision Contrat x Garantie, comme son nom l'indique, elle nous intéresse, car elle nous permet de récupérer toutes les informations relatives à chacune des garanties de notre base. Pour pouvoir filtrer sur le code produit, nous joignons les BV Garantie et Contrat afin de récupérer le code produit dans la BV Contrat. Nous filtrons ensuite sur le code produit pour conserver les produits qui nous intéressent puis sur le code garantie pour conserver uniquement les garanties Indemnités Journalières. Les informations que nous souhaitons extraire de la BV Garantie sont :

- L'état de la garantie (en cours, résilié, éteint, sans effet, sinistré, résilié sans valeur) pour chaque garantie,
- La date d'effet de la garantie,
- Les franchises accident, maladie et hospitalisation liées à chacune des garanties,
- La date du dernier changement de statut de la garantie et,
- La date d'effet de la garantie.

La base garantie a une vision ligne, alors que nous souhaitons extraire cette information avec une vision colonne, exemple avec la garantie 1 :

Garantie 1	Etat de la garantie 1	Franchise maladie	Franchise accident	Franchise hospitalisation	Date début effet garantie 1

Pour cela, nous filtrons la base Garantie sur la garantie 1 (pour nous, il s'agit des IJ toutes causes) puis nous créons une sous-base comportant l'état de la garantie, son libellé et le numéro du contrat. Nous joignons ensuite la sous-base à notre BV CRM par numéro de contrat. Nous réalisons cette opération pour chacune des garanties.

ii) Enrichissement de la base CRM avec la base contrat de VANDA

La base contrat de VANDA nous permet de récupérer de précieuses informations relatives aux différents contrats dont le niveau de sélection médicale. Elle est traitée comme suit avant la réalisation de la jointure :

- Transformation de tous les « null » en « Non » pour les variables de sélection médicale.
- Création d'une variable de niveau de sélection médicale :
  1. Niveau de sélection médicale 1
  2. Niveau de sélection médicale 2
  3. Niveau de sélection médicale 3
  4. Niveau de sélection médicale 4
  5. Niveau de sélection médicale 5

  Tous les contrats doivent passer par de la sélection médicale (plus le niveau de sélection médicale est élevé, de 1 à 5 et plus la sélection médicale est poussée). Il existe des contrats où tous les questionnaires sont "null" : cela concerne majoritairement un produit qui a subi une migration de systèmes, et certaines informations n'ont pas été correctement rattrapées.

  De plus, normalement, le niveau de sélection médicale 5 seul n'existe pas, il est accompagné du Niveau de sélection médicale 3. Néanmoins, il existe des contrats qui ont seulement le niveau de sélection médicale 5. Il s'agit uniquement de contrats Produit 4.
- Suppression des doublons par rapport au numéro de contrat (en théorie, la BV Contrat a une vision contrat, néanmoins, il existe quelques cas de doublons très marginaux). Il n'y a aucun doublon sur notre périmètre Prévoyance Individuelle.

La base Contrat a une vision contrat. La jointure avec la BV CRM se fait simplement par numéro de contrat.

iii) Enrichissement de la base CRM avec la base personne de VANDA

Cette étape nous permet d'ajouter à nos données les informations relatives aux assurés de notre base. On suit les étapes suivantes pour la réaliser :

- Nous appliquons un filtre sur les rôles « Assuré » et « Souscripteur - Assuré »,
- Et nous supprimons les doublons : au cas où, car dans des cas très marginaux, il peut y avoir des doublons pour un même numéro de contrat. Il peut s'agir de lignes concernant maris et femmes, mais ce problème de doublons ne concerne aucun des contrats présents dans notre base, et ne semble pas concerner les contrats de Prévoyance Individuelle. Il s'agit très probablement des contrats d'assurance sur 2 têtes.

La Base Personne ayant une vision Contrat. La jointure se fait simplement par numéro de contrat.

iv) Enrichissement de la base CRM avec la base surprime de VANDA

La base surprime nous permet de récupérer les informations sur le fait qu'un assuré paie oui ou non une surprime. Pour cette base, nous procédons comme suit :

— Nous créons une sous-base avec tous les contrats ayant au moins une garantie surprimée. Nous n'avons pris que les garanties incapacité, vu que c'est notre périmètre d'étude. Nous créons une variable "TOP Surprime" dans cette sous-base prenant la modalité « OUI » pour tous les contrats présents. Au moment de la jointure des bases, pour les contrats n'ayant pas de surprime, ce TOP Surprime prendra la modalité « null », que l'on remplacera par un « NON ».

— Nous faisons de même pour le "TOP Réduction". De cette manière, la BV Surprime a une vision contrat. La jointure se fait simplement par numéro de contrat.

v) Enrichissement de la base CRM avec la base sinistre de VANDA

La jointure fut beaucoup plus compliquée pour cette base. En effet, nous avons dû nous limiter à la jointure sur les dossiers acceptés de notre base CRM (soit 18 000 dossiers) et avons dû créer de nouvelles clés primaires pour réaliser la jointure, car les numéros de sinistre de VANDA et CRM étaient différents et le numéro de contrat de l'assuré seul ne suffit pas pour retrouver un sinistre. Nous avons donc dû procéder selon plusieurs étapes qui seront plus détaillées ci-dessous et faire la jointure petit à petit en utilisant différentes clés primaires.

1. Afin de réaliser la jointure, nous créons une clé à partir du numéro de contrat, de la date de survenance du sinistre et du « numéro ordre extension ». Ce dernier représente l'ordre du dossier au sein du sinistre, il vaut 0 pour les dossiers d'ouverture et 1 pour le 1e dossier de prolongation, 2 pour le 2e dossier de prolongation, 3 pour le 3e dossier de prolongation et ainsi de suite. Nous ne l'avons pas dans la base CRM donc on le crée et comme on est seulement sur les ouvertures, on le met à 0 pour tous les dossiers. Nous obtenons une jointure pour 70% de la base (soit 12 600 dossiers).
2. Pour la deuxième jointure, nous créons une clé primaire à partir du numéro de contrat, la date de début de l'AI et le numéro ordre extension sinistre. Nous arrivons à joindre plus de 80% des dossiers acceptés (soit 14 400 dossiers).
3. Cette fois, pour la troisième jointure, pour réaliser notre clé primaire, nous utilisons :
   - Le numéro de contrat,
   - La date de fin de l'AI,
   - Le numéro ordre extension sinistre.

Nous arrivons à passer de 80% de dossiers acceptés joints à 90% (soit 16 200 dossiers).

vi) Enrichissement de la base CRM avec la deuxième base garantie de VANDA

Nous joignons cette base à la base contrat de VANDA afin d'avoir en plus du code garantie, le code produit et donc de pouvoir filtrer sur les produits et garanties qui nous intéressent.

Les variables qui nous intéressent dans cette base sont :

— « VALEUR » qui contient le montant de la garantie (pas forcément une garantie IJ) et

— « MESURE » qui contient la devise (points pour Francs ou Euro).

Nous filtrons la base obtenue (après jointure entre les bases contrat VANDA et stock NT garantie) sur le « type – rubrique » « Sinistre » et sur les « libellés nature » qui comportent l'intitulé indemnités journalières. Nous corrigeons la variable "VALEUR" en prenant en compte la devise.
Nous filtrons la deuxième base garantie sur la garantie 1 (pour nous, il s'agit des IJ toutes causes) puis nous créons une sous-base comportant le numéro de contrat et le montant d'IJ. Nous joignons ensuite la sous-base à notre Base obtenue après l'ajout de la Base sinistre par numéro de sinistre. Nous réalisons cette opération pour chacune des garanties afin d'obtenir tous les montants d'IJ pour elles toutes.
Nous arrivons à avoir ces montants seulement pour les dossiers acceptés pour lesquels la jointure entre notre base CRM (à laquelle nous avons ajouté les bases contrat, surprime, personne et garantie) et notre base sinistre de VANDA a fonctionné (en effet ce sont les seuls dossiers pour lesquels nous avons pu avoir le numéro de sinistre VANDA).

vii) Récapitulatif des jointures réalisées

Nous retrouverons dans l'annexe A les différentes étapes schématisées des différentes jointures réalisées.

3.3.2 Tests de conformité et création des variables finales

Une fois toutes les jointures réalisées, nous réalisons des comparaisons entre les variables CRM et leurs équivalents VANDA.

i) Tests de conformité

Nous nous plaçons sur la totalité de nos dossiers 18 000 acceptés (les 16 200 dossiers acceptés pour lesquels la jointure a fonctionné et les 1 800 dossiers restants pour lesquels la jointure avec la BV sinistre de VANDA n'a pas fonctionné) et nous calculons le nombre de dossiers pour lesquels nous avons égalité entre la valeur que nous trouvons sur VANDA et la valeur obtenue dans CRM pour avoir une idée de la fiabilité des données que nous avons obtenues sur CRM pour les dossiers refusés. Nous obtenons les taux de correspondance suivants :

Variables	**Assiette**	**Nombre de dossiers avec égalité entre CRM et VANDA**	**Nombre de dossiers avec variable CRM non-nulle**	**Nombre de dossiers avec variable VANDA non-nulle**	**Taux de correspondance**
Date de début d'AI	17 800	12 816	15 000	16 200	72%
Date de fin d'AI	17 200	12 040	14 000	16 200	70%
Date de début d'hospi	3 700	851	1 000	3 600	23%
Date de fin d'hospi	3 700	555	1 000	3 600	15%
Date de survenance	16 700	13 193	15 500	16 200	79%
Montant d'IJ garantie 1	15 700	14 130	14 500	15 200	90%
Montant d'IJ garantie 2	8 300	7 470	8 000	7 000	90%
Montant d'IJ garantie 3	4 200	3 780	4 000	3 500	90%
Montant d'IJ garantie 4	4 800	3 600	4 500	4 000	75%
Montant d'IJ garantie 5	475	437	450	400	92%

TABLE 3.3 – Taux de correspondance entre CRM et VANDA(1/4)

Variables	**Assiette**	**Nombre de dossiers avec égalité entre CRM et VANDA**	**Nombre de dossiers avec variable CRM non-nulle**	**Nombre de dossiers avec variable VANDA non-nulle**	**Taux de correspondance**
Montant d'IJ garantie 6	411	370	400	370	90%
Montant d'IJ garantie 7	75	68	70	50	90%
Franchise accident garantie 1	11 380	10 242	10 200	11 000	90%
Franchise accident garantie 2	7 200	6 480	6 200	7 000	90%
Franchise accident garantie 3	2 150	1 935	1 600	2 000	90%
Franchise accident garantie 4	4 200	3 570	3 750	4 000	85%
Franchise accident garantie 5	475	413	420	450	87%
Franchise accident garantie 6	420	378	380	400	90%
Franchise accident garantie 7	63	50	57	60	80%
Durée max d'indemnisation garantie 4	4 000	3 120	3 900	3 500	78%

TABLE 3.4 – Taux de correspondance entre CRM et VANDA(2/4)

Variables	**Assiette**	**Nombre de dossiers avec égalité entre CRM et VANDA**	**Nombre de dossiers avec variable CRM non-nulle**	**Nombre de dossiers avec variable VANDA non-nulle**	**Taux de correspondance**
Franchise hospi garantie 1	11 220	10 098	10 197	11 000	90%
Franchise hospi garantie 2	7 300	6 570	6 325	7 000	90%
Franchise hospi garantie 3	2 200	1 804	1 926	2 000	82%
Franchise hospi garantie 4	4 200	3 486	3 775	4 000	83%
Franchise hospi garantie 5	470	404	400	450	86%
Franchise hospi garantie 6	411	350	350	400	85%
Franchise hospi garantie 7	61	50	52	60	82%
Franchise maladie garantie 1	11 500	9 200	10 250	11 000	80%
Franchise maladie garantie 2	7 100	5 964	6 560	7 000	84%
Franchise maladie garantie 3	2 100	1 806	1 830	2 000	86%
Franchise maladie garantie 4	4 300	3 827	3 700	4 000	89%

TABLE 3.5 – Taux de correspondance entre CRM et VANDA(3/4)

Variables	**Assiette**	**Nombre de dossiers avec égalité entre CRM et VANDA**	**Nombre de dossiers avec variable CRM non-nulle**	**Nombre de dossiers avec variable VANDA non-nulle**	**Taux de correspondance**
Franchise maladie garantie 5	451	389	415	435	86%
Franchise maladie garantie 6	441	397	419	430	90%
Durée max d'indemnisation garantie 1	11 725	9 145	11 000	11 500	78%
Durée max d'indemnisation garantie 2	7 040	5 773	6 540	7 000	82%
Durée max d'indemnisation garantie 3	2 120	1 802	1 920	2 000	85%
Durée max d'indemnisation garantie 5	430	370	422	430	86%
Durée max d'indemnisation garantie 6	440	374	390	420	85%
Durée max d'indemnisation garantie 7	68	51	60	50	75%
Top hospitalisation	3 828	957	2 505	3 600	25%
Cause sinistre	16 750	13 400	13 800	16 200	80%

TABLE 3.6 – Taux de correspondance entre CRM et VANDA(4/4)

Dans les tableaux ci-dessus, chaque colonne correspond à l'élément suivant :

- Variable : comme son nom l'indique, c'est dans cette colonne qu'on met les variables qui nous intéressent,
- Assiette : cette colonne correspond au nombre de dossiers pour lesquels nous avons une valeur non-nulle dans CRM et/ou VANDA.
- Nombre de dossiers avec égalité entre CRM et VANDA : cette colonne contient le nombre de dossiers pour lesquels la valeur obtenue dans CRM est égale à la valeur que nous avons dans VANDA,
- Nombre de dossiers avec variable CRM non-nulle : elle contient le nombre de dossiers pour lesquels nous disposons de l'information dans CRM,
- Nombre de dossiers avec variable VANDA non-nulle : elle contient le nombre de dossiers pour lesquels l'information est disponible dans VANDA,
- Le taux de correspondance : cette colonne contient le quotient entre la 3 colonne et la 2 colonne.

ii) Variables finales

Les données disponibles dans VANDA étant plus fiables que les données CRM, nous construisons les variables finales comme suit :

- Cause sinistre : nous prenons la valeur de la variable CAUSE SINISTRE de VANDA si elle est renseignée et sinon nous prenons la valeur de la variable CAUSE SINISTRE CRM, puis nous créons 4 nouvelles modalités (maladie, accident, maladie + hospitalisation, accident + hospitalisation) en prenant en compte l'hospitalisation,
- Pour les autres variables (date de début de l'AI, date de fin de l'AI, date de survenance du sinistre, ...), nous prenons la valeur de VANDA quand elle existe, sinon, nous prenons celle de CRM.

Pour les variables que nous avons pu trouver seulement sur VANDA ou seulement sur CRM nous n'avons pas pu effectuer de comparaison et nous avons gardé comme variable finale la variable de départ.

3.4 Vérification de la complétude des variables finales

Avant de vérifier la complétude des variables finales, nous décidons d'apporter quelques modifications à notre base de données :

- Nous décidons dans un premier temps de remplacer les valeurs manquantes de la variable "Date de début d'AI" par les valeurs de la variable "Date de survenance du sinistre" si celle-ci existe, car nous avons constaté que les valeurs de ces variables sont très proches. De la même manière, nous remplaçons les valeurs manquantes de la variable "Date de survenance du sinistre" par les valeurs de la variable "Date de début de l'AI".

— Puis, après avoir remarqué pour certains contrats ne sont rattachés à aucune garantie, nous créons une variable "Top garantie" qui dit si oui ou non, nous avons au moins une garantie pour un contrat donné.

— Ensuite, nous supprimons les dossiers pour lesquels nous n'avons toujours pas la date de début de l'arrêt, il ne nous reste plus que 22 100 dossiers (dont 17 500 acceptés et 4 500 refusés, cela représente 18 000 contrats) et,

— Enfin, nous créons une nouvelle variable nombre de sinistres qui nous donne le nombre de sinistres pour un contrat donné à date de la nouvelle déclaration.

Nous calculons les taux de complétudes des variables finales créées pour nos 22 100 dossiers et nous obtenons les résultats contenus dans les tableaux suivants :

Variables	**Assiette**	**Nombre de dossiers pour lesquelles la variable est non-nulle**	**Taux de complétude**
Date de fin d'AI	22 100	17 000	77%
Cause du sinistre	22 100	20 200	91%
Top hospitalisation	22 100	22 100	100%
Date de début d'hospi	3 600	3 600	100%
Date de fin d'hospi	3 600	3 595	99%
Montant d'IJ garantie 1	19 200	18 700	97%
Montant d'IJ garantie 2	10 000	8 240	82%
Montant d'IJ garantie 3	3 000	2 980	99%
Montant d'IJ garantie 4	5 900	5 005	85%
Montant d'IJ garantie 5	475	460	97%
Montant d'IJ garantie 6	457	450	98%
Montant d'IJ garantie 7	70	68	97%
Franchise accident garantie 1	19 200	19 000	99%
Franchise accident garantie 2	10 000	9 225	92%
Franchise accident garantie 3	3 000	2 989	99%
Franchise accident garantie 4	5 900	5 600	95%
Franchise accident garantie 5	475	475	100%
Franchise accident garantie 6	457	450	98%
Franchise accident garantie 7	70	69	99%

TABLE 3.7 – Taux de complétude des variables finales (1/3)

Variables	**Assiette**	**Nombre de dossiers pour lesquelles la variable est non-nulle**	**Taux de complétude**
Niveau de sélection médicale	22 100	21 000	95%
Top fumeur	22 100	20 000	90%
Franchise maladie garantie 1	19 200	19 050	99%
Franchise maladie garantie 2	10 000	9 000	90%
Franchise maladie garantie 3	3 000	2 930	98%
Franchise maladie garantie 4	5 900	5 444	92%
Franchise maladie garantie 5	475	465	98%
Franchise maladie garantie 6	457	453	99%
Franchise hospitalisation garantie 1	19 200	18 240	95%
Franchise hospitalisation garantie 2	10 000	9 000	90%
Franchise hospitalisation garantie 3	3 000	2 790	93%
Franchise hospitalisation garantie 4	5 900	5 310	90%
Franchise hospitalisation garantie 5	475	451	95%
Franchise hospitalisation garantie 6	457	434	95%
Franchise hospitalisation garantie 7	70	65	93%

TABLE 3.8 – Taux de complétude des variables finales (2/3)

Variables	**Assiette**	**Nombre de dossiers pour lesquelles la variable est non-nulle**	**Taux de complétude**
Durée max d'indemnisation garantie 1	19 200	17 664	92%
Durée max d'indemnisation garantie 2	10 000	8 500	85%
Durée max d'indemnisation garantie 3	3 000	2 700	90%
Durée max d'indemnisation garantie 4	5 900	5 251	89%
Durée max d'indemnisation garantie 5	475	441	93%
Durée max d'indemnisation garantie 6	457	420	92%
Durée max d'indemnisation garantie 7	70	63	90%

TABLE 3.9 – Taux de complétude des variables finales (3/3)

Les colonnes de ces tableaux correspondent à :

— Variable : le nom de la variable considérée,

— Assiette : c'est le nombre de dossiers pour lesquels on doit avoir l'information dans la base de données,

— Nombre de dossiers pour lesquels la variable est non-nulle : cette colonne, comme son nom l'indique correspond au nombre de dossiers pour lesquels l'information est disponible dans la BV (les dossiers pour lesquels nous n'avons pas une valeur nulle),

— Taux de complétude : cette colonne contient le quotient entre la 2 colonne et la première colonne.

Pour les variables "Sexe", "Classe Tarifaire", "Réseau de distribution", "Date de survenance", "Date de déclaration du sinistre", "Top surprime" et "Date de début de l'arrêt", nous n'avons aucune valeur manquante et donc un taux de complétude égal à 100%.

Nous constatons la présence de valeurs manquantes pour certaines variables de notre jeu de données, dans la suite, nous verrons comment les traiter.

3.5 Traitement des valeurs manquantes pour les variables finales

3.5.1 Mise en place de modèles de machine learning et sélection des meilleurs modèles

Nous décidons, afin de conserver le maximum de données possibles, de compléter des lignes avec valeurs manquantes plutôt que de les supprimer.
Dans un premier temps, nous utilisons des méthodes simples pour certaines variables. Nous retrouvons ci-dessous les variables concernées par cette première approche et les méthodes utilisées.

— Pour la date de fin d'hospitalisation, nous avons 5 dossiers pour lesquels la date de début d'hospitalisation est renseignée, mais pas la date de fin. Dans ce cas, nous décidons de supprimer les dates de début d'hospitalisation ce qui permet d'avoir une symétrie d'informations, nous modifions aussi la valeur du top hospitalisation aussi pour ces dossiers en le mettant en "NON".

— et pour les montants d'IJ pour toutes les garanties, nous décidons de remplacer les valeurs manquantes par les valeurs médianes (nous avons avant cette étape vérifiée qu'il n'existait pas de valeurs aberrantes pour ces variables).

Dans un second temps, nous utilisons des méthodes de machine learning pour compléter les variables : "Date de fin de l'AI", "Cause du sinistre", "niveau de sélection médicale", "Top fumeur", "Franchise accident, maladie et hospitalisation" et enfin "Durée maximale d'indemnisation" (seulement pour les garanties 1, 2 et 4 ; pour les autres, nous remplaçons les valeurs manquantes par la valeur la plus fréquente, car nous n'avons pas assez de données pour mettre en place des modèles de machine learning assez robustes).

Nous expliquerons dans la suite comment nous procédons pour compléter ces valeurs manquantes.

Date de fin d'AI

Ne pouvant pas calibrer un modèle de machine learning sur des variables de type **date**, afin de pouvoir mener à bien la complétion de la variable "Date de fin d'AI", nous décidons de créer une nouvelle variable "Durée de l'AI" qui donne la durée entre la date de début de l'arrêt et sa date de fin.

Nous décidons d'utiliser les modèles de forêt aléatoire et de réseau de neurones, car ils sont relativement faciles à mettre en place et ne nécessitent pas d'étude de corrélation de variables par exemple. En effet, nous pouvons penser de manière intuitive qu'il y a de nombreuses variables corrélées dans notre base telles que :

— La cause du sinistre et le top hospitalisation,

— Ou encore la date de survenance du sinistre et le nombre de sinistres.

- **Partition des données**
  Nous allons partitionner les données en une base contenant les 5 100 dossiers sans date de fin d'AI, une seconde qui nous permettra de calculer les scores du modèle grâce à la **validation croisée**.
- **Mise en place du modèle de forêt aléatoire**
  Nous avions tenté d'utiliser les outils *grid search*, *randomized search* et *optuna* mais les algorithmes mettaient beaucoup trop de temps à tourner. Pour le *grid search* par exemple, nous avions décidé de faire varier les hyper-paramètres de la manière suivante :
  - Le nombre d'arbres de notre modèle est entre 100 et 2000,
  - La profondeur de nos arbres comprise entre 5 et 50,
  - Et enfin le nombre d'éléments minimum contenus dans chaque nœud compris entre 2 et 20.

  Cependant, le code avait tourné pendant plus de 48 heures et ne donnait toujours pas de résultats, nous avons donc décidé de sélectionner les meilleurs hyperparamètres en tâtonnant et nous avons observé que le modèle avec le plus faible RMSE qui est égal à 53 est obtenu pour les valeurs suivantes :
  - Nombre d'arbres = 500,
  - Profondeur des arbres = 8,
  - Et enfin nombre d'éléments minimum dans chaque nœud = 2.
- **Mise en place du modèle de réseau de neurones**
  Nous gardons les mêmes partitions du jeu de données, encore une fois les outils permettant d'obtenir les meilleurs hyperparamètres ne donnant rien, nous tâtonnons pour les obtenir (nous partirons sur des neurones ayant 2 à 3 couches cachées).
  Nous obtenons comme meilleurs hyperparamètres les éléments suivants :
  - Nombre de couches cachées = 3,
  - le nombre de neurones dans chaque couche cachée = (35,60,40),
  - la fonction d'activation = relu,
  - et enfin le paramètre de régularisation L2 = 0,001.

  Et ces éléments nous permettent d'avoir le RMSE suivant : 55.

Nous constatons que le meilleur modèle est le modèle de forêt aléatoire. C'est donc lui que nous utilisons pour compléter les valeurs manquantes pour la durée d'AI, ensuite il ne nous reste plus qu'à ajouter la durée prédite à la date de début d'AI associé à un dossier donné pour obtenir la date de fin de l'AI.

Cause du sinistre

Nous procédons comme précédemment pour le découpage du jeu de données et pour la sélection des meilleurs modèles, sauf que nous choisissons l'accuracy comme métrique

de comparaison des modèles. En effet, vu que notre variable à prédire a 4 modalités, c'est la métrique qui nous semble être la plus adaptée. Le meilleur modèle pour cette variable est la forêt aléatoire avec les hyper-paramètres suivants :

- Nombre d'arbres = 1000,
- Profondeur des arbres = 30,
- Et enfin nombre d'éléments minimum dans chaque nœud = 10.

Et avec ces paramètres, nous obtenons un accuracy de 85% contre environ 81% pour le meilleur modèle de réseau de neurones pour les paramètres ci-dessous :

- Nombre de couches cachées = 3,
- le nombre de neurones dans chaque couche cachée = (20,60,40),
- la fonction d'activation = relu,
- la méthode d'optimisation des poids = adam,
- et enfin le maximum d'itérations pour la convergence = 200.

Après cette étape, la variable est complète.

Niveau de sélection médicale

Nous avons 22 100 dossiers, ce qui représente 18 000 contrats au total. Et sur ces 18 000 contrats, nous en avons 400 pour lesquels le niveau de sélection médicale n'est pas renseigné (soit 1 100 dossiers). Nous utilisons les méthodes de découpage des données et de sélection des meilleurs hyper-paramètres pour cette variable et nous prédisons les valeurs du niveau de sélection médicale pour ces 500 contrats. Nous décidons aussi de définir l'accuracy comme métrique de sélection du modèle le plus judicieux, car comme dans le cas de la variable "cause du sinistre", nous avons jusqu'à 5 modalités pour le "niveau de sélection médicale".
L'accuracy le plus élevé est obtenu pour le modèle de forêt aléatoire avec les mêmes hyper-paramètres que ceux obtenus pour la cause du sinistre (Nombre d'arbres =1000, Profondeur des arbres=30 et Nombre d'éléments minimum dans chaque nœud =10). Pour ce modèle, nous obtenons un accuracy de 82% contre 78% pour le meilleur modèle de réseau de neurones avec deux couches cachées et les hyper-paramètres suivants :

- Nombre de couches cachées = 2,
- le nombre de neurones dans chaque couche cachée = (20,64),
- la fonction d'activation = relu,
- la méthode d'optimisation des poids = adam,
- et enfin le maximum d'itérations pour la convergence = 200.

Nous arrivons à compléter cette variable à 100%.

Top fumeur

Nous avons sur les 18 000 contrats de notre base de données, 830 pour lesquels les valeurs du top fumeur sont manquantes (soit 2 100 dossiers). Ce sont tous des contrats rattachés aux produits 3 (80%) et 4 (20%). Nous procédons comme pour le remplissage des valeurs manquantes de la variable "Niveau de sélection médicale" sauf que nous utilisons l'AUC pour comparer les modèles.
Nous obtenons l'AUC de plus élevé pour tous les modèles de forêt aléatoire (environ 0,63) pour les mêmes hyperparamètres obtenus le modèle retenu pour le niveau de sélection médicale.
Pour le réseau de neurones, nous obtenons aussi les mêmes hyperparamètres les plus adaptés que ceux obtenus pour le niveau de sélection médicale, mais avec un AUC d'environ 0,615.

Nous arrivons bien à compléter cette variable.

Les variables franchise accident, maladie, hospitalisation, la durée maximale d'indemnisation pour les garanties 1, 2 et 4

Ces variables étant considérées comme des variables qualitatives (en effet, bien qu'elles soient numériques, nous n'en disposons que d'un nombre bien définit), nous utilisons les mêmes méthodes que pour la variable "Top fumeur" et nous arrivons à toutes les compléter.

3.5.2 Récapitulatif des variables complétées et des méthodes utlisées

Ci-dessous, des tableaux permettant de faire un récapitulatif des variables que nous avons complétées ainsi que les méthodes que nous avons utilisées :

Variables	**Nombre de dossiers ou contrats avec VM**	**Modèle le plus adapté**	**Hyper-paramètres**	**Métrique retenue et sa valeur**
Date de fin d'AI	5 100 dossiers	Forêt aléatoire	(500,8,2)	RMSE=53
Cause du sinistre	1 900 dossiers	Forêt aléatoire	(1000,30,10)	Accuracy=85%
Niveau de sélection médicale	400 contrats	Forêt aléatoire	(1000,30,10)	Accuracy=82%
Top fumeur	830 contrats	Forêt aléatoire	(1000,30,10)	AUC=0,63
Franchise accident garantie 1	80 contrats	Forêt aléatoire	(1000,30,10)	Accuracy=89%
Franchise maladie garantie 1	67 contrats	forêt aléatoire	(1000,20,10)	Accuracy=84%
Franchise hospitalisation garantie 1	230 contrats	Forêt aléatoire	(1000,20,5)	Accuracy=85%
Durée maximale d'indemnisation garantie 1	465 contrats	Forêt aléatoire	(1000,30,10)	Accuracy=87%

TABLE 3.10 – Complétion des valeurs manquantes (1/2)

Variables	**Nombre de dossiers ou contrats avec VM**	**Modèle le plus adapté**	**Hyper-paramètres**	**Métrique retenue et sa valeur**
Franchise accident garantie 2	200 contrats	forêt aléatoire	(1000,30,10)	Accuracy=89%
Franchise maladie garantie 2	372 contrats	Forêt aléatoire	(900,30,8)	Accuracy=79%
Franchise hospitalisation garantie 2	372 contrats	Forêt aléatoire	(1000,30,10)	Accuracy=80%
Durée maximale d'indemnisation garantie 2	420 contrats	Forêt aléatoire	(1000,30,10)	Accuracy=89%
Franchise accident garantie 4	97 contrats	Forêt aléatoire	(800,30,10)	Accuracy=89%
Franchise maladie garantie 4	107 contrats	Forêt aléatoire	(1000,30,5)	Accuracy=87%
Franchise hospitalisation garantie 4	120 contrats	Forêt aléatoire	(1000,30,10)	Accuracy=87%
Durée maximale d'indemnisation garantie 4	139 contrats	Forêt aléatoire	(1000,20,10)	Accuracy=84%

TABLE 3.11 – Complétion des valeurs manquantes (2/2)

3.6 Vérification de la cohérence de nos prédictions

Une fois nos données complétées, nous réalisons différents tests de logiques pour vérifier si les valeurs obtenues sont cohérentes.
Pour se faire, nous décidons de réaliser une première analyse sur nos données.

— Nous vérifions dans un premier temps si pour les dossiers refusés pour lesquels nous avons complété les durées de sinistre manquantes, les motifs statut sont cohérents à ce que nous observons dans les données.
Comme nous l'avions brièvement évoqué dans le tableau D.1, le motif statut peut prendre diverses valeurs dont les plus récurrentes sont :

- Franchise qui dit si le dossier est refusé pour cause de franchise atteinte,
- Pas de garantie souscrite qui dit si le dossier est refusé parce que rattaché à aucune garantie,
- Délai d'attente qui dit que le dossier est refusé parce que le délai d'attente est atteint,
- Et enfin l'exclusion qui dit si la pathologie ayant entrainé l'arrêt fait partie des clauses d'exclusion.

Sur les 5 100 dossiers pour lesquels nous complétons les valeurs manquantes de la durée du sinistre, 3 700 sont refusés et pour ces dossiers refusés, nous faisons les observations suivantes :

- Sur 3 700 dossiers, 1 500 le sont pour motif de franchise atteinte et en comparant les durées de sinistre aux franchises de la première garantie applicable (on sélectionne bien la franchise en fonction de la cause du sinistre), on se rend compte que pour une grande partie (70% soit 1 050 dossiers) la durée du sinistre est supérieure la durée de la Franchise appliquée.
- Cette incohérence peut aussi être due au fait que nous avons complété les valeurs manquantes pour les franchises et les causes de sinistre.
- Nous ne pouvons donc pas dire que la complétion des valeurs manquantes pour la durée du sinistre, sa cause et les différentes franchises fonctionne bien.

— Ensuite, en comparant les reporting des produits 1 et 2 pour mars 2023 à nos données, nous réalisons que nous n'avons pas exactement les mêmes répartitions de contrats de part et d'autre pour chacun des niveaux de sélection médicale.
En effet, nous avons des résultats assez similaires pour les niveaux 1 et 2 (différence de 2,25% pour le niveau 1 et 0,5% seulement pour le niveau 2) dans le reporting et nos données, mais des écarts assez importants pour les niveaux 3 et 4 (5% pour le niveau 3 et 4,25% pour le niveau 4).
Ces écarts ne sont pas surprenants au vu du volume important de dossiers pour lesquels nous n'avions pas les niveaux de sélection médicale, mais cela nous envoie néanmoins à nous questionner sur la qualité de nos méthodes de complétion des valeurs manquantes pour cette variable.

Tout ceci nous pousse à croire que la complétion de nos valeurs manquantes a créé un biais dans nos données et donc nous pensons que des modèles mis en place sur de telles données ne seront pas assez robustes et ne donneront pas des résultats intéressants.
Nous décidons donc dans la suite de notre étude de ne garder que les lignes complètes de notre base de données (celles pour lesquelles nous n'avons pas eu à compléter les valeurs manquantes)

Chapitre 4

Analyse des données

4.1 Analyse descriptive des données

Nous réalisons l'analyse de nos données sur les lignes qui étaient complètes et pour lesquelles nous n'avions pas eu besoin de compléter les valeurs manquantes, nous avons donc 17 000 dossiers (dont 13 039 acceptés et 3 961 refusés) pour 14 000 contrats dans notre base.

4.1.1 Analyse descriptive univariée

Nous réalisons les analyses sur les variables suivantes :

— **Les variables liées au sinistre** : Il s'agit des informations qui sont récupérées seulement une fois le sinistre survenu et déclaré. Nous pouvons citer par exemple la cause du sinistre, la durée de l'AI, l'âge à la survenance du sinistre, le top déclaration tardive qui vaut "OUI" si le sinistre est déclaré au moins 30 jours après sa survenance et "NON" sinon,... Pour tous les graphiques ci-dessous, nous nous intéressons à la répartition des sinistres (nous pouvons remplacer le terme dossier par sinistre ici, car nous travaillons uniquement sur les ouvertures d'arrêt de travail) pour une variable donnée :

- **Durée de l'AI**

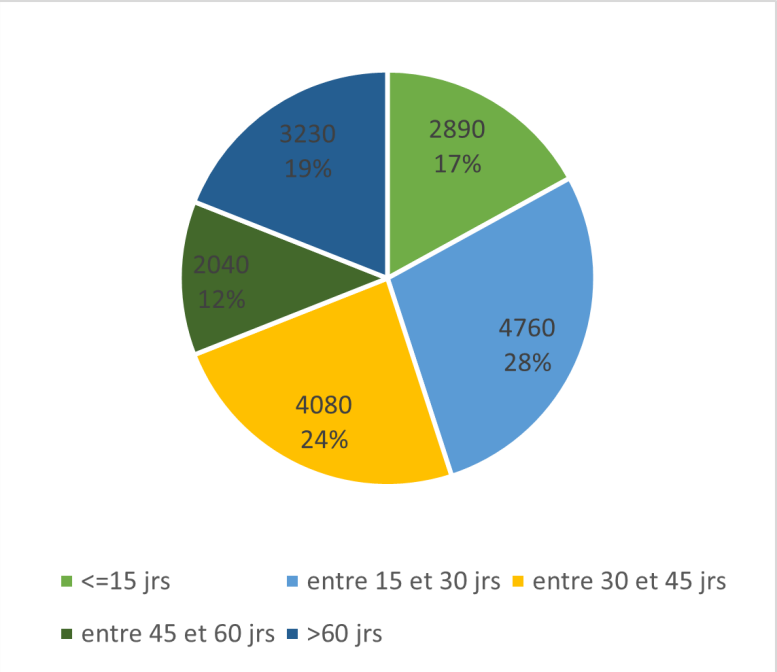


FIGURE 4.1 – Nombre de sinistres par rapport à la durée de l'arrêt

Nous remarquons que nous avons plus de sinistres avec une durée "moyenne" c'est-à-dire entre 30 et 45 jours dans notre base.

- **Cause du sinistre**

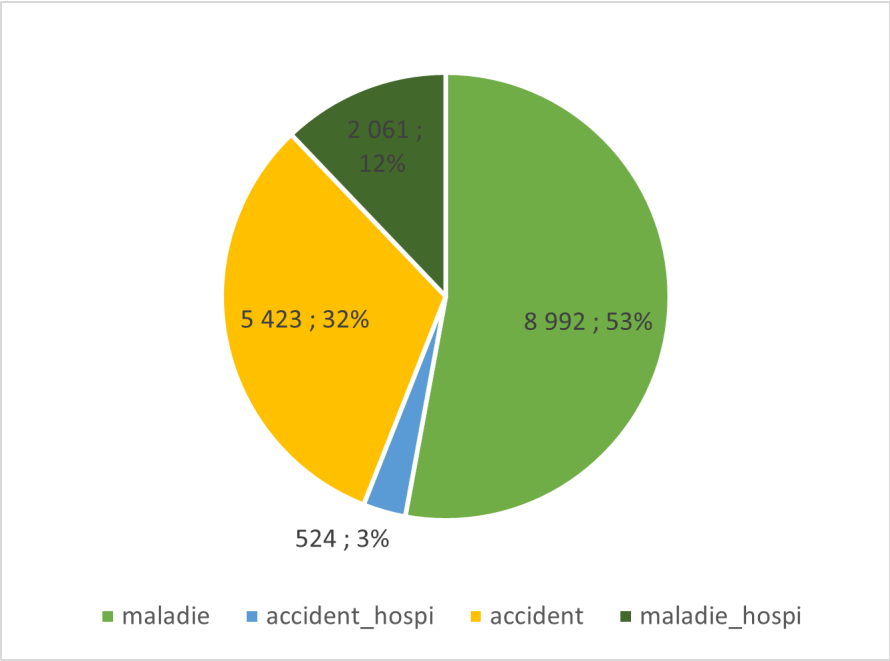


FIGURE 4.2 – Nombre de sinistres par rapport à la cause du sinistre

Nous remarquons que la cause la plus fréquente dans la base est la "maladie".

- **Âge de survenance du sinistre**

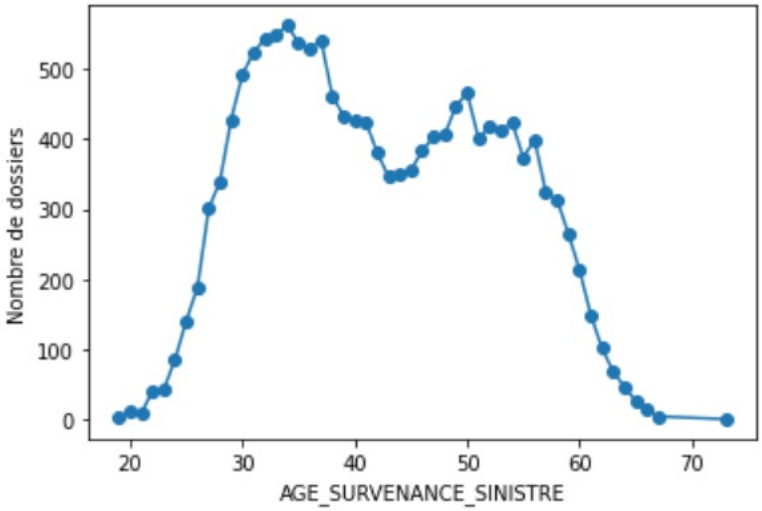


FIGURE 4.3 – Nombre de sinistres par rapport à l'âge de la survenance

Nous remarquons que nous avons plus de sinistres survenus entre 30 et 56 ans, mais nous observons une baisse de sinistres dans la quarantaine, cela pourrait être dû à des grossesses pathologiques entre 30 et 40 ans, car l'âge moyen de l'accouchement en France est de 31 ans (nous pourrons mieux observer ce phénomène quand nous représenterons le graphique sur le croisement entre l'âge de survenance et le sexe). Mais, cela pourrait aussi s'expliquer par le fait que les individus ayant entre 30 et 56 ans au moment du sinistre sont les plus présents dans notre base, ce qui est normal vu que c'est dans cette tranche d'âge que les gens sont le plus en activité.

- **Durée de déclaration**

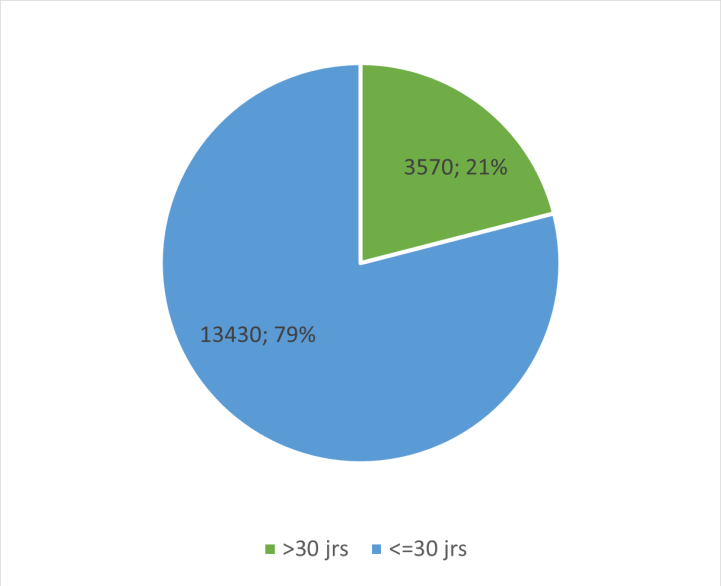


FIGURE 4.4 – Nombre de sinistres par rapport à la durée entre la survenance et la déclaration

Nous remarquons que nous avons peu de sinistres pour lesquels la durée entre la date de survenance du sinistre et sa date de déclaration dépasse 30 jours.

— **Les variables liées aux contrats** : ce sont toutes les informations dont nous disposons dès que le contrat est souscrit, nous pouvons citer par exemple dans cette catégorie : le sexe de l'assuré, le top surprime, la sélection médicale …. Nous retrouverons ci-dessous les graphiques des différentes analyses mises en place pour quelques-unes de ces variables.

- **Niveau de sélection médicale**

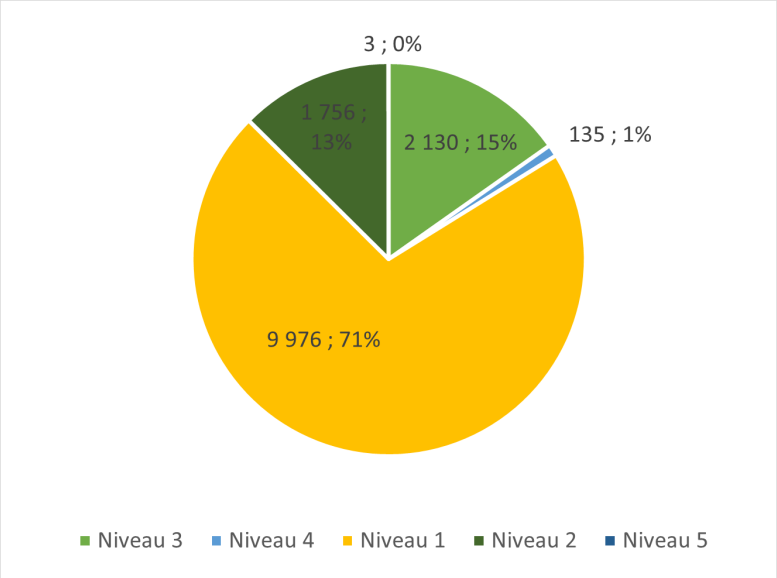


FIGURE 4.5 – Nombre de contrats pour chaque modalité de la sélection médicale

Le niveau de sélection médicale le plus fréquent dans la base est le "Niveau 1", ce qui est normal vu qu'il s'agit du questionnaire médical renseigné à minima.

- **Top surprime**

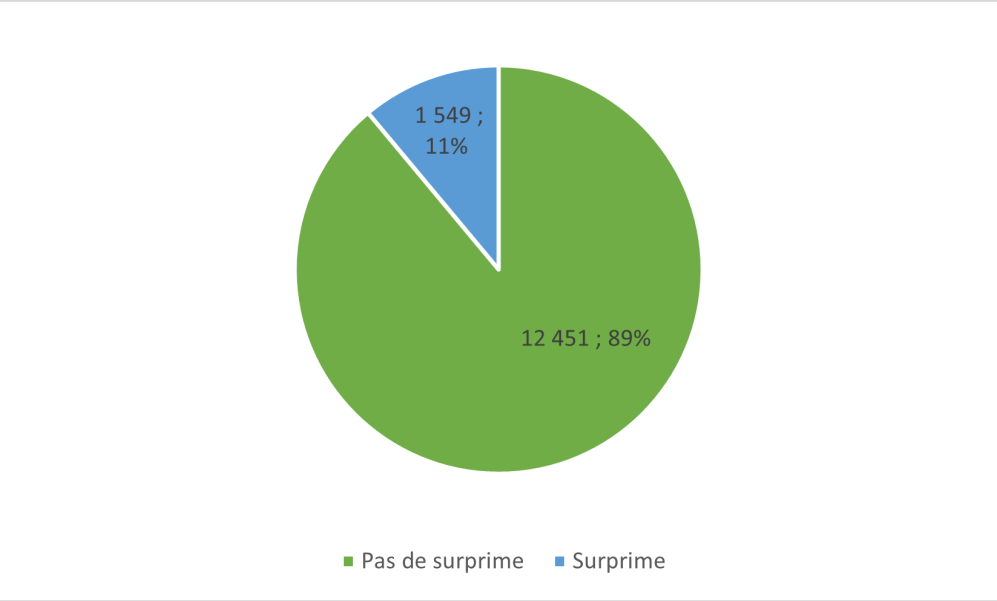


FIGURE 4.6 – Nombre de contrats pour chaque modalité de la surprime

Nous avons plus de contrats sans surprime dans notre base, cela est cohérent avec ce que l'on observe en portefeuille.

- **Top fumeur**

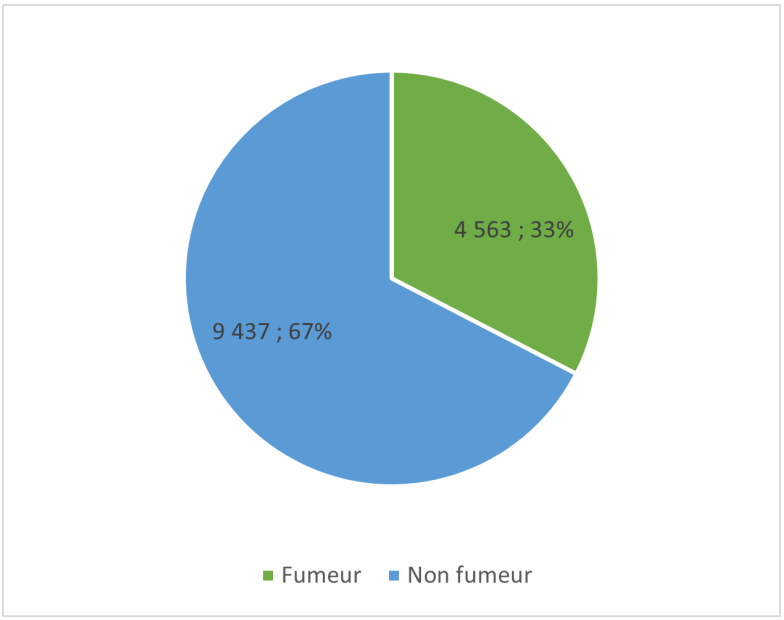


FIGURE 4.7 – Nombre de contrats pour chaque modalité du top fumeur

Nous observons plus de contrats pour lesquels l'assuré est non-fumeur que de contrats pour lesquels il est fumeur.

- **Nombre de sinistres**

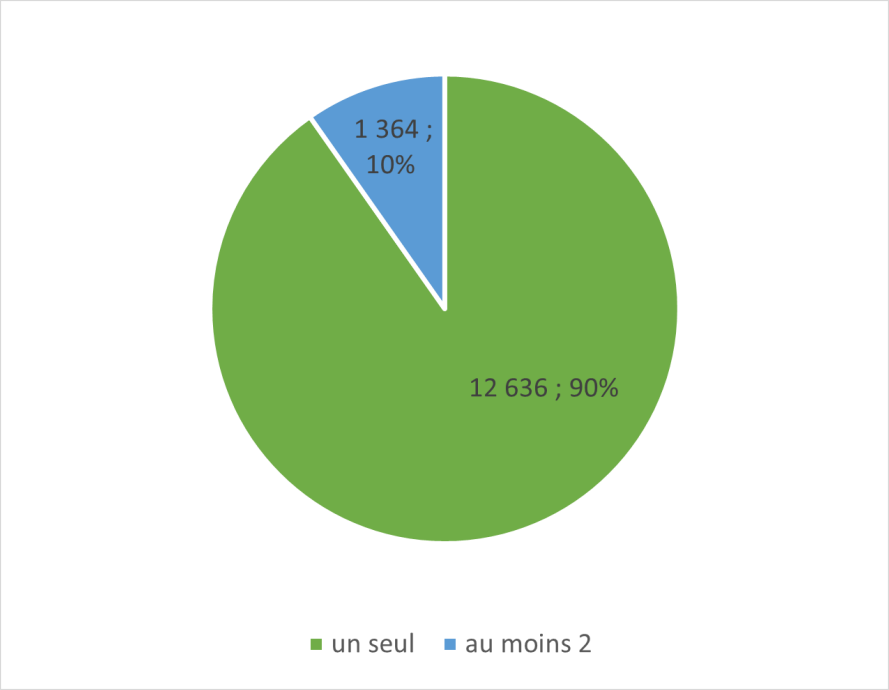


FIGURE 4.8 – Nombre de contrats avec un seul sinistre et avec plus d'un sinistre

Nous remarquons que la plupart des contrats ont seulement 1 sinistre dans la base de données.

4.1.2 Analyse descriptive bivariée

Dans cette partie, nous procédons à l'analyse des données en se basant sur les interactions entre certaines variables. Nous avons choisi les variables dont les interactions nous semblaient intéressantes à étudier pour la suite de notre étude.

Nous obtenons les graphiques ci-dessous en étudiant les interactions entre certaines variables liées au sinistre et d'autres liées au contrat.

- **Cause du sinistre et sexe de l'assuré**

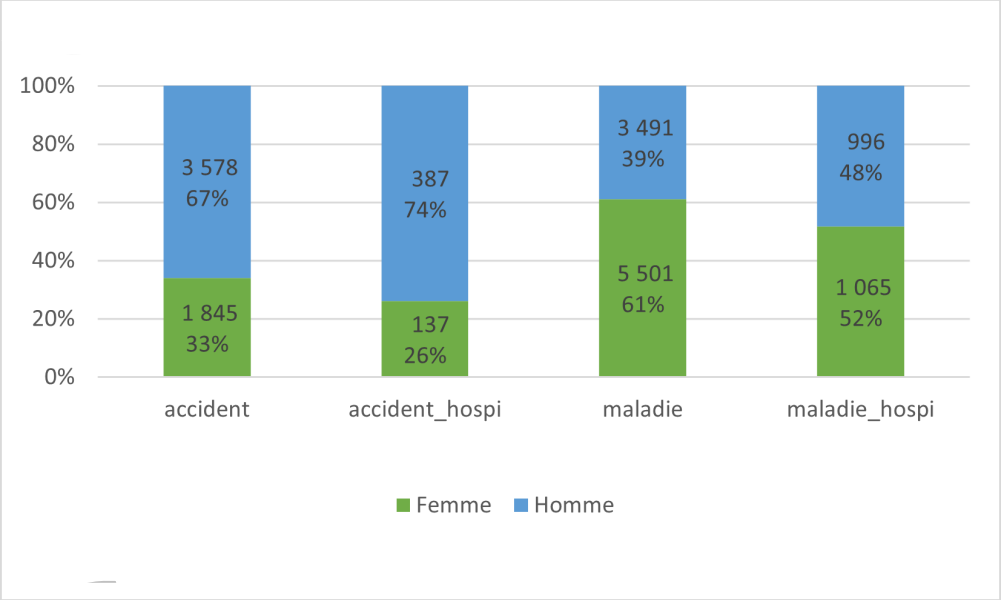


FIGURE 4.9 – Nombre de sinistres par cause de sinistre et sexe

Nous remarquons que contrairement aux causes de sinistre "maladie" et "maladie et hospitalisation" pour lesquelles la répartition homme/femme est plutôt homogène, pour les causes "accident" et "accident et hospitalisation", nous avons pratiquement 3 fois plus d'hommes que de femmes.

- **Âge de survenance et sexe de l'assuré**

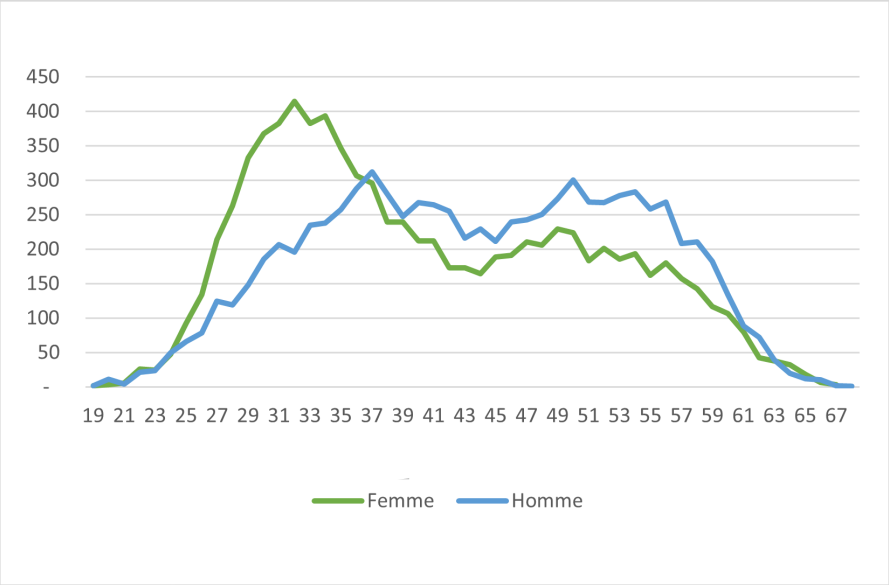


FIGURE 4.10 – Nombre de sinistres par âge de survenance et sexe

Nous remarquons que contrairement à la courbe de l'âge pour les hommes, celle des femmes croît très tôt, entre 27 et 33 ans, et ensuite connait une baisse pendant la quarantaine. Cela pourrait être dû à la présence de nombreux sinistres pour grossesse pathologique dans notre base et est cohérent avec l'âge moyen de l'accouchement pour les femmes en France qui est de 31 ans [16].

- **Âge de survenance et cause du sinistre**

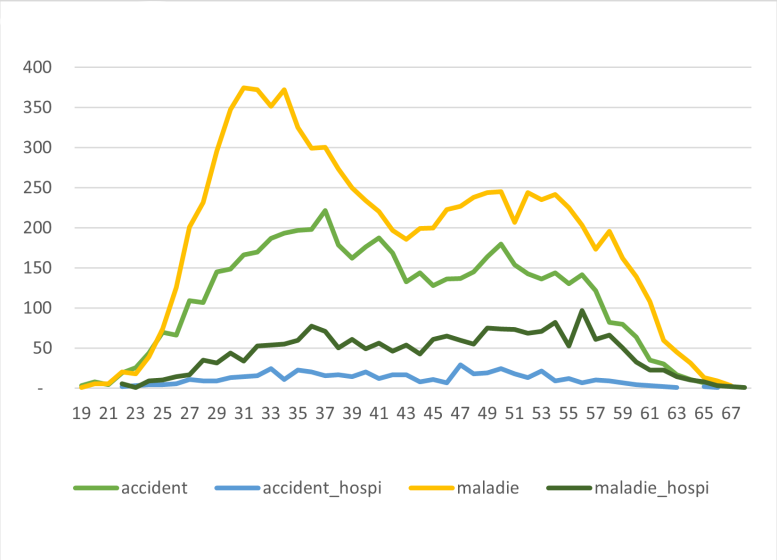


FIGURE 4.11 – Nombre de sinistres par âge de survenance et cause

Nous remarquons que la combinaison de ces deux variables ne donne pas plus d'informations supplémentaire, car le volume de données dont nous disposons pour les causes "accident hospitalisation" et "maladie hospitalisation" n'est pas assez significatif pour faire des conclusions justes.

- **Cause de sinistre et top fumeur**

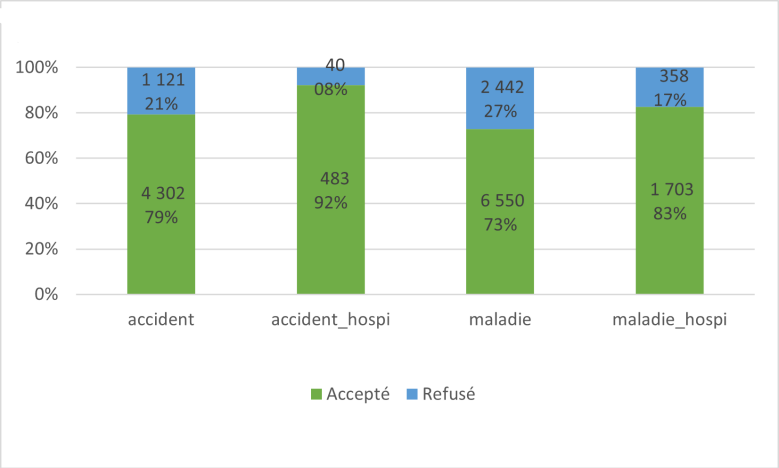


FIGURE 4.12 – Nombre de sinistres par cause de sinistre et top fumeur

Nous observons une homogénéité de la répartition fumeur/non-fumeur pour toutes les modalités de la variable cause sinistre. Nous avons chaque fois plus de non-fumeurs que de fumeur et ce quelle que soit la cause du sinistre.

Les résultats de ces premières analyses nous paraissent cohérents pour la suite de notre étude.

4.1.3 Analyse par rapport à l'OUTCOME

Nous décidons après l'analyse descriptive globale de notre base de se pencher sur l'analyse par rapport à la variable sur laquelle nous comptons réaliser nos modèles, c'est-à-dire l'OUTCOME. Ces analyses ont été mises en place dans le but de mieux observer les effets de certaines variables sur notre variable à expliquer.
Nous obtenons pour quelques-unes des variables étudiées dans le cadre des analyses univariées et bivariées globales les graphiques suivants :

- **Durée du sinistre par rapport à la décision.**

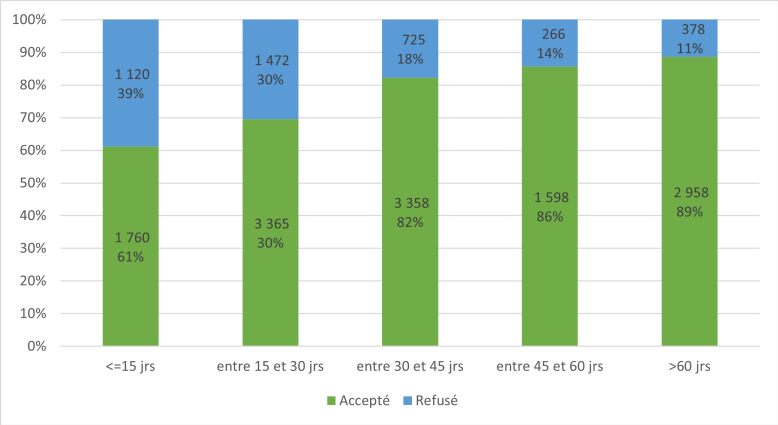


FIGURE 4.13 – Nombre de sinistres par durée et décision

Nous remarquons que la plupart des dossiers refusés concernent soit des demandes pour des sinistres très courts (qui durent moins de 15 jours) soit des sinistres de durée moyenne (entre 30 et 45 jours).

- **Niveau de sélection médicale par rapport à la décision**

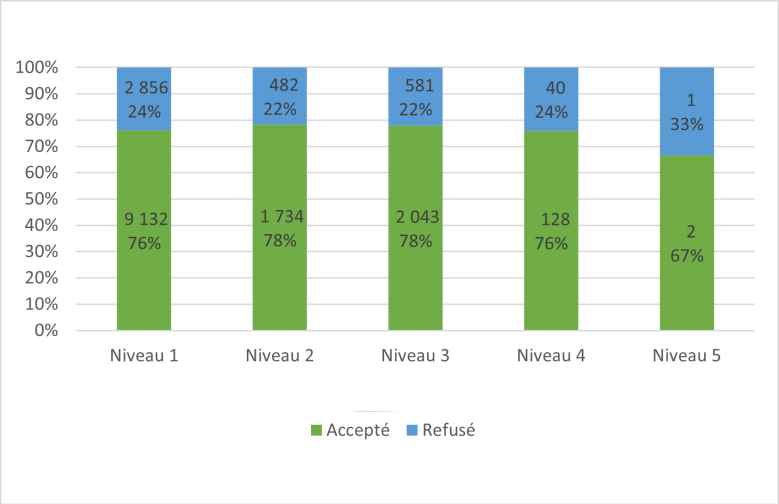


FIGURE 4.14 – Nombre de sinistres par sélection médicale et décision

Nous remarquons que nous avons un peu plus de dossiers refusés pour les niveaux de sélection médicale "Niveau de sélection médicale 1" et "Niveau de sélection médicale 4". Cependant, vu le très faible volume de données pour le niveau 4, nous ne pouvons pas vraiment apporter une interprétation du résultat de l'analyse pour lui.

- **Cause du sinistre par rapport à la décision**

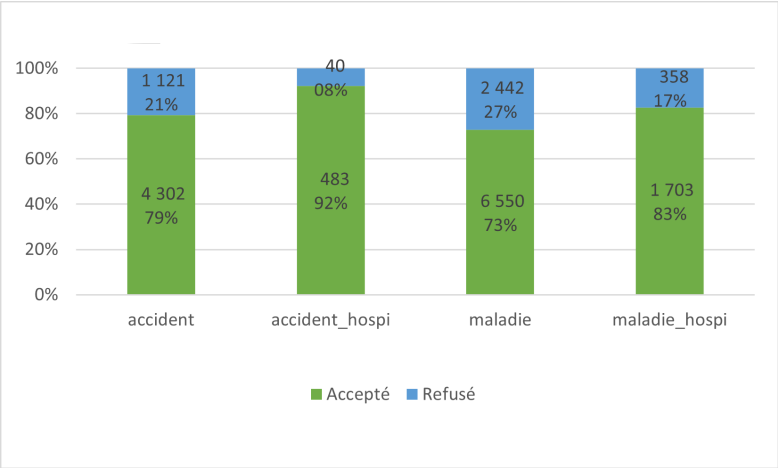


FIGURE 4.15 – Nombre de sinistres par cause et décision

Nous avons beaucoup plus de dossiers refusés pour la cause maladie que pour les autres causes de sinistres à la fois en termes de quantité et de proportion.

- **Réseau de distribution et décision**

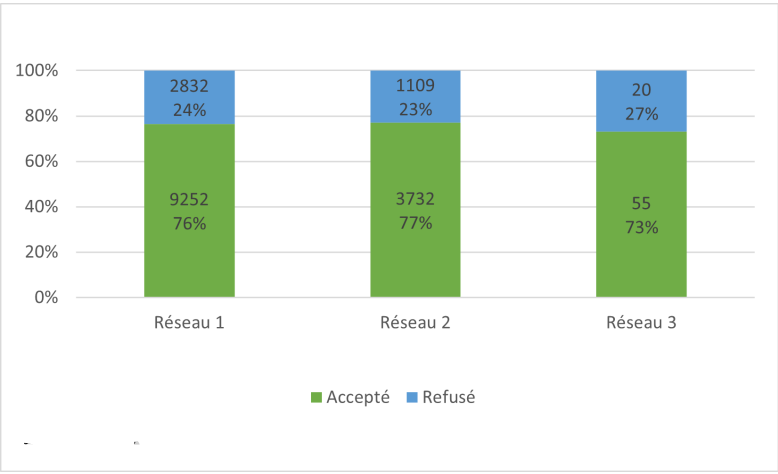


FIGURE 4.16 – Nombre de sinistres par réseau de distribution et décision

Nous observons plus de dossiers refusés pour les réseaux réseau 1 et réseau 3, cela peut s'expliquer par le fait que le réseau 1 soit le réseau de distribution avec le plus de contrats (sur notre périmètre d'étude) et donc de dossiers dans la base.

Il est important de noter que la réalisation d'une analyse descriptive sur la variable "PROFESSION" par rapport à l'OUTCOME aurait été très intéressante pour notre

étude, cependant, cela ne semblait pas raisonnable au vu de la quantité astronomique de modalités dont nous disposions dans cette variable.

Aussi, pour la suite de notre étude, nous décidons de "déplier" notre base de données, en effet avoir une vision une ligne par dossiers fait que nous avons toujours des valeurs manquantes dans nos données pour les variables liées aux garanties. Cela est dû au fait que tous les assurés n'ont pas toutes les garanties.

Nous passons donc à une vision une ligne par dossier et garantie souscrite, ce qui fait que nous avons maintenant plusieurs fois le même dossier dans notre base de données selon le nombre de garanties auquel il est rattaché.

Nous avons alors un jeu de données contenant 33 000 lignes dont 25 000 concernent des dossiers acceptés et 8 000 des dossiers refusés.

4.1.4 Apport des résultats des analyses

L'analyse descriptive des données nous a permis de voir que nos données pouvaient être déséquilibrées par rapport à certaines variables. Cela est par exemple le cas pour la cause du sinistre car nous remarquons que plus de la moitié des arrêts de notre base sont causés par une maladie alors que seulement 03% sont causés par un accident suivi d'une hospitalisation. C'est aussi le cas de la variable niveau de sélection médicale pour laquelle nous pouvons constater que plus de 71% des contrats de notre base sont de niveau 1 contre seulement 1% pour le niveau 4.

Nous devrons donc être vigilants au moment de l'interprétation des résultats de nos modèles car des déséquilibres trop importants peuvent biaiser les résultats obtenus.

L'analyse nous a aussi permis d'avoir des pistes sur les profils de dossiers qui seraient plus susceptibles d'être acceptés (et inversement). En effet, l'analyse de la durée du sinistre par rapport à l'OUTCOME a montré que plus le sinistre est court et plus il est fréquent qu'il soit refusé.

4.2 Étude de corrélation entre les variables

Afin de mener à bien la mise en place de nos modèles, nous décidons d'étudier les corrélations entre les différentes variables. En effet, avoir des variables avec une corrélation élevée dans un modèle de machine learning peut conduire à l'obtention de résultats complètement erronés et dans certains cas même cela peut être à l'origine de la non-convergence de nos modèles, surtout pour les modèles linéaires généralisés.

Il est vrai que certains modèles sont plus sensibles aux corrélations entre les variables explicatives que d'autres, mais vu que nous utiliserons 3 modèles dans notre étude et dans le but d'obtenir les résultats les plus justes possibles, c'est une étape qui nous paraît indispensable.

4.2.1 Corrélation de Pearson

i) Explications théoriques de la corrélation de Pearson

Ce coefficient de corrélation [17] permet de mesurer la relation de proportionnalité entre les valeurs prises par deux variables dans une base, qu'elles soient de type qualitatif ou quantitatif.
Supposons que nous avons une base de données appelée X qui contient n lignes, chaque ligne représentant un individu et m colonnes, chacune représentant une variable.
Pour deux variables X_i et X_j de notre base, le coefficient de corrélation de Pearson se calcule comme suit :

$$\sigma_{i,j} = \frac{COV(X_{.,i}, X_{.,j})}{\sqrt{VAR(X_{.,i}) \times VAR(X_{.,j})}}$$

Où :

$$COV(X_{.,i}, X_{.,j}) = \frac{1}{n} \sum_{h=0}^{n} (x_{h,i} - \bar{x}_i)(x_{h,j} - \bar{x}_j)$$

Avec :

— $\bar{x}_i$ et $\bar{x}_j$ qui représentent les moyennes respectives des variables $X_{.,i}$ et $X_{.,j}$,

— Et $x_{h,i}$ et $x_{h,j}$ qui représentent les observations de ces variables.

Ce coefficient de corrélation varie entre -1 et 1 et son interprétation est relativement simple : plus il est proche de 1 en valeur absolue et plus la corrélation entre les deux variables est importante.

— S'il est proche de 1 en valeur absolue et positif, nous disons que les variables sont très corrélées positivement,

— Sinon nous disons qu'elles sont très corrélées négativement.

ii) Application à nos données

Nous calculons les coefficients de corrélation entre nos différentes variables (nous gardons uniquement les coefficients de corrélation supérieurs à 0,5) et nous obtenons les valeurs renseignées dans les tableaux ci-dessous :

Variable 1	Variable 2	Corrélation
CAUSE_SINISTRE_VF_accident	CAUSE_SINISTRE_VF_maladie	-0,72
CAUSE_SINISTRE_VF_maladie_hospi	TOP_HOSPI_VF_NON	-0,88
CAUSE_SINISTRE_VF_maladie_hospi	TOP_HOSPI_VF_OUI	0,88
CAUSE_SINISTRE_VF_maladie_hospi	DUREE_HOSPI_0	-0,62
CAUSE_SINISTRE_VF_maladie_hospi	DUREE_HOSPI__15	0,59
TOP_HOSPI_VF_NON	TOP_HOSPI_VF_OUI	-1,00
TOP_HOSPI_VF_NON	DUREE_HOSPI_0	0,70
TOP_HOSPI_VF_NON	DUREE_HOSPI__15	-0,67
TOP_HOSPI_VF_OUI	DUREE_HOSPI_0	-0,70
TOP_HOSPI_VF_OUI	DUREE_HOSPI__15	0,67
NBRE_SINISTRES_VF_plusieurs	NBRE_SINISTRES_VF_un_seul	-1,00
TOP_REFUS_AV_DECLA_NON	TOP_REFUS_AV_DECLA_OUI	-1,00
MONTANT_IJ_VF_100_300	MONTANT_IJ_VF__100	-0,99
FR_ACC_VF_3	FR_ACC_VF_365	-0,53
FR_ACC_VF_3	FR_MAL_VF_365,0	-0,53
FR_ACC_VF_3	FR_HOSPI_VF_3	0,68
FR_ACC_VF_3	FR_HOSPI_VF_365	-0,53
FR_ACC_VF_3	DUREE_MAX_INDM_VF_365	0,54
FR_ACC_VF_3	DUREE_MAX_INDM_VF_1095	-0,54
FR_ACC_VF_3	GARANTIE_Garantie1	-0,51
FR_ACC_VF_15	FR_HOSPI_VF_15	1,00
FR_ACC_VF_30	FR_MAL_VF_30,0	0,53
FR_ACC_VF_30	FR_HOSPI_VF_30	1,00
FR_ACC_VF_90	FR_MAL_VF_90,0	1,00
FR_ACC_VF_90	FR_HOSPI_VF_90	1,00
FR_ACC_VF_365	FR_MAL_VF_365,0	1,00
FR_ACC_VF_365	FR_HOSPI_VF_365	1,00
FR_ACC_VF_365	DUREE_MAX_INDM_VF_365	-0,53
FR_ACC_VF_365	DUREE_MAX_INDM_VF_1095	0,53
FR_ACC_VF_365	GARANTIE_Garantie1	0,67
FR_MAL_VF_30,0	FR_HOSPI_VF_30	0,53
FR_MAL_VF_90,0	FR_HOSPI_VF_90	1,00
FR_MAL_VF_365,0	FR_HOSPI_VF_365	1,00
FR_MAL_VF_365,0	DUREE_MAX_INDM_VF_365	-0,53
FR_MAL_VF_365,0	DUREE_MAX_INDM_VF_1095	0,53
FR_MAL_VF_365,0	GARANTIE_Garantie1	0,67
FR_HOSPI_VF_365	DUREE_MAX_INDM_VF_365	-0,53

TABLE 4.1 – Tableau de corrélation des variables (1/3)

Variable 1	Variable 2	Corrélation
FR_HOSPI_VF_365	DUREE_MAX_INDM_VF_1095	0,53
FR_HOSPI_VF_365	GARANTIE_Garantie1	0,67
DUREE_MAX_INDM_VF_365	DUREE_MAX_INDM_VF_1095	-1,00
DUREE_MAX_INDM_VF_365	GARANTIE_Garantie1	-0,74
DUREE_MAX_INDM_VF_365	GARANTIE_Garantie2	0,76
DUREE_MAX_INDM_VF_1095	GARANTIE_Garantie1	0,74
DUREE_MAX_INDM_VF_1095	GARANTIE_Garantie2	-0,76
LIBELLE_ETAT_GARANTIE_En_cours	LIBELLE_ETAT_GARANTIE_Résilié	-1,00
LIBELLE_ETAT_GARANTIE_En_cours	LIBELLE_ETAT_EXO_INCAP_Résilié	-0,61
LIBELLE_ETAT_GARANTIE_En_cours	STATUT_CONTRAT_VF_en_cours	0,90
LIBELLE_ETAT_GARANTIE_En_cours	STATUT_CONTRAT_VF_résilié	-0,90
LIBELLE_ETAT_GARANTIE_Eteint	LIBELLE_ETAT_EXO_INCAP_Eteint	0,56
LIBELLE_ETAT_GARANTIE_Résilié	LIBELLE_ETAT_EXO_INCAP_Résilié	0,61
LIBELLE_ETAT_GARANTIE_Résilié	STATUT_CONTRAT_VF_en_cours	-0,90
LIBELLE_ETAT_GARANTIE_Résilié	STATUT_CONTRAT_VF_résilié	0,90
LIBELLE_ETAT_GARANTIE_Sans_effet	LIBELLE_ETAT_EXO_INCAP_Sans_effet	0,80
LIBELLE_ETAT_GARANTIE_Sans_effet	STATUT_CONTRAT_VF_sans_effet	0,72
LIBELLE_ETAT_GARANTIE_Sinistré	LIBELLE_ETAT_EXO_INCAP_Sinistré	0,76
LIBELLE_ETAT_GARANTIE_Sinistré	STATUT_CONTRAT_VF_Sinistré	0,53
LIBELLE_ETAT_EXO_INCAP_Résilié	STATUT_CONTRAT_VF_en_cours	-0,58
LIBELLE_ETAT_EXO_INCAP_Résilié	STATUT_CONTRAT_VF_résilié	0,58
LIBELLE_ETAT_EXO_INCAP_Sans_effet	STATUT_CONTRAT_VF_sans_effet	0,58
STATUT_CONTRAT_VF_en_cours	STATUT_CONTRAT_VF_résilié	-0,99
SEXE_VF_Féminin	SEXE_VF_Masculin	-1,00
TOP_SURPRIME_VF_NON	TOP_SURPRIME_VF_OUI	-1,00
NSM_NIVEAU1	NSM_NIVEAU3	-0,66
NSM_NIVEAU1	NSM_NIVEAU2	-0,59
TOP_FUMEUR_VF_1_oui	TOP_FUMEUR_VF_2_non	-1,00
RESEAU_VF_RESEAU1	RESEAU_VF_RESEAU2	-0,99
MOTIF_STATUT_VF_Clôturée_Règlement	MOTIF_STATUT_VF_Franchise	-0,54
TYPE_FLUX_VF_Espace_Client	TYPE_FLUX_VF_Hors_Espace_Client	-1,00

TABLE 4.2 – Tableau de corrélation des variables (2/3)

Variable 1	Variable 2	Corrélation
PRESENCE_ENFANT_VF_NON	PRESENCE_ENFANT_VF_OUI	-1,00
TOP_NOUVELLE_ADHESION_1_oui	TOP_NOUVELLE_ADHESION_2_non	-1,00
CLAUSE_VF_NON	CLAUSE_VF_OUI	-1,00
DUREE_HOSPI_0	DUREE_HOSPI__15	-0,96
AGE_SURVENANCE_SINISTRE_33_61	AGE_SURVENANCE_SINISTRE__33	-0,93
AGE_SURVENANCE_SINISTRE_33_61	AGE_FIN_AI_33_61	0,98
AGE_SURVENANCE_SINISTRE_33_61	AGE_FIN_AI__33	-0,91
AGE_SURVENANCE_SINISTRE_33_61	AGE_CONTRAT_33_61	0,69
AGE_SURVENANCE_SINISTRE_33_61	AGE_CONTRAT__33	-0,67
AGE_SURVENANCE_SINISTRE_62_	AGE_FIN_AI_62_	0,96
AGE_SURVENANCE_SINISTRE_VI	AGE_FIN_AI_VI	0,84
AGE_SURVENANCE_SINISTRE__33	AGE_FIN_AI_33_61	-0,91
AGE_SURVENANCE_SINISTRE__33	AGE_FIN_AI__33	0,98
AGE_SURVENANCE_SINISTRE__33	AGE_CONTRAT_33_61	-0,74
AGE_SURVENANCE_SINISTRE__33	AGE_CONTRAT__33	0,75
AGE_FIN_AI_33_61	AGE_FIN_AI__33	-0,92
AGE_FIN_AI_33_61	AGE_CONTRAT_33_61	0,67
AGE_FIN_AI_33_61	AGE_CONTRAT__33	-0,66
AGE_FIN_AI__33	AGE_CONTRAT_33_61	-0,73
AGE_FIN_AI__33	AGE_CONTRAT__33	0,73
AGE_CONTRAT_33_61	AGE_CONTRAT__33	-0,99
DUREE_EFFET_GAR_SURV_365_	DUREE_EFFET_GAR_SURV_90_365	-0,83
DUREE_EFFET_GAR_SURV_365_	DUREE_EFFET_AI_365_	0,94
DUREE_EFFET_GAR_SURV_365_	DUREE_EFFET_AI_90_365	-0,83
DUREE_EFFET_GAR_SURV_90_365	DUREE_EFFET_AI_365_	-0,86
DUREE_EFFET_GAR_SURV_90_365	DUREE_EFFET_AI_90_365	0,98
DUREE_EFFET_GAR_SURV__90	DUREE_EFFET_AI__90	0,80
PRODUIT_PRODUIT1	PRODUIT_PRODUIT2	-0,95
DUREE_EFFET_AI_365_	DUREE_EFFET_AI_90_365	-0,88
GARANTIE_Garantie1	GARANTIE_Garantie2	-0,60

TABLE 4.3 – Tableau de corrélation des variables (3/3)

Cette étude de corrélation nous a permis de constater que comme on pouvait s'en douter, certaines variables de notre base sont très corrélées.
Il s'agit de :

— Les variables liées aux garanties, les franchises accident, maladie et hospitalisation ont des coefficients de corrélation très élevés entre elles,

— Les variables top hospitalisation et durée d'hospitalisation aussi sont très corrélés, il faut aussi noter que la variable cause du sinistre est très corrélée à la variable top hospitalisation,

Nous verrons dans la suite comment nous choisirons les variables que nous maintiendrons dans le modèle de départ et celles que nous supprimerons.

4.2.2 Test de Khi 2

i) Explications théoriques du test de Khi 2

Il existe différents types de tests de Khi2 selon ce que l'on souhaite observer. Nous avons par exemple les tests de khi 2 de l'indépendance, de distribution et de l'homogénéité.
Dans cette étude, on parlera uniquement du test d'indépendance de khi 2 et nous comparerons les résultats obtenus aux résultats des tests de corrélation vus plus tôt.

Ce test est important, car il permet de décider si les variables qualitatives pour lesquelles il est réalisé sont probablement liées ou non. Nous disons que les variables sont indépendantes s'il n'existe aucun lien statistique entre elles, autrement dit si la connaissance de l'une d'entre elle ne permet pas de se prononcer sur l'autre. Pour le mettre en place, on se base sur deux hypothèses :

— L'hypothèse H_0 pour laquelle nous affirmons que les deux variables sont indépendantes,

— L'hypothèse H_1 qui est l'hypothèse alternative ou contraire.

Le test renvoie trois valeurs importantes (le test statistique de khi 2, le degré de liberté, df et le niveau de signification, α).
Dans cette étude, nous nous intéresserons plus aux valeurs du test statistique de khi 2 et du niveau de significativité.
Ces deux éléments s'interprètent de la manière suivante :

— Le test statistique de Khi 2 : il permet de mesurer la différence entre les fréquences observées et les fréquences obtenues si les variables étaient bien indépendantes. Plus il est élevé et plus la différence entre les données observées et les données attendues est aussi élevée, ce qui veut dire que les variables sur lesquels le test a été réalisé ne sont **indépendantes**.

— Le niveau de signification α : c'est le critère permettant de déterminer si les résultats de notre test de Khi 2 sont significatifs ou pas. Le plus souvent, on le prend égal à 5% et on le compare à la p-value comme suit :

  - Si la valeur de la p-value obtenue est inférieure à α alors l'hypothèse nulle est rejetée, ce qui veut dire que la relation entre les deux variables testées est significative et donc qu'elles ne sont certainement pas indépendantes.
  - Dans le cas contraire, on ne peut pas conclure que les variables sont dépendantes.

ii) Application à nos données

L'application du test de Khi 2 à nos données nous donne des résultats que l'on pourra retrouver dans l'annexe B cependant, nous en afficherons un aperçu ci-dessous (il est important de noter que nous gardons uniquement les p-values qui sont supérieures au niveau $\alpha = 0,05$) :

Variable 1	Variable 2	Chi-squared	P-value
CAUSE_SINISTRE_VF	TOP_REFUS_AV_DECLA	6,04	0,11
CAUSE_SINISTRE_VF	TOP_NOUVELLE_ADHESION	11,07	0,09
TOP_HOSPI_VF	NBRE_SINISTRES_VF	2,39	0,12
TOP_HOSPI_VF	TOP_REFUS_AV_DECLA	0,20	0,66
TOP_HOSPI_VF	MONTANT_IJ_VF	0,39	0,82
TOP_HOSPI_VF	TOP_FUMEUR_VF	1,83	0,18
TOP_HOSPI_VF	TOP_NOUVELLE_ADHESION	1,51	0,47
NBRE_SINISTRES_VF	TOP_HOSPI_VF	2,39	0,12
NBRE_SINISTRES_VF	DUREE_MAX_INDM_VF	0,52	0,47
NBRE_SINISTRES_VF	TOP_SURPRIME_VF	0,01	0,91
NBRE_SINISTRES_VF	TOP_FUMEUR_VF	0,06	0,81
NBRE_SINISTRES_VF	PRESENCE_ENFANT_VF	0,02	0,89
TOP_REFUS_AV_DECLA	CAUSE_SINISTRE_VF	6,04	0,11
TOP_REFUS_AV_DECLA	TOP_HOSPI_VF	0,20	0,66
TOP_REFUS_AV_DECLA	MONTANT_IJ_VF	0,38	0,83
TOP_REFUS_AV_DECLA	FR_ACC_VF	10,30	0,17
TOP_REFUS_AV_DECLA	FR_MAL_VF	9,94	0,27
TOP_REFUS_AV_DECLA	FR_HOSPI_VF	13,64	0,09
TOP_REFUS_AV_DECLA	DUREE_MAX_INDM_VF	1,01	0,31
TOP_REFUS_AV_DECLA	TOP_SURPRIME_VF	0,99	0,32
TOP_REFUS_AV_DECLA	CODE_REGION_PORTEFEUILLE_VF	6,42	0,27
TOP_REFUS_AV_DECLA	PRESENCE_ENFANT_VF	2,45	0,12
TOP_REFUS_AV_DECLA	TOP_NOUVELLE_ADHESION	4,06	0,13
TOP_REFUS_AV_DECLA	DUREE_HOSPI	8,97	0,11
TOP_REFUS_AV_DECLA	AGE_CONTRAT	3,23	0,36
TOP_REFUS_AV_DECLA	GARANTIE	1,92	0,75
MONTANT_IJ_VF	TOP_HOSPI_VF	0,39	0,82
MONTANT_IJ_VF	TOP_REFUS_AV_DECLA	0,38	0,83

TABLE 4.4 – Aperçu des résultats du test de Khi 2

Les résultats de ce test viennent confirmer les résultats obtenus pour l'étude des corrélations des variables.

4.3 Variables à conserver pour nos modèles

En prenant en compte les résultats obtenus à partir des différentes analyses réalisées ainsi que des différents tests de corrélation mis en place, nous décidons de conserver les variables citées ci-dessous pour constituer nos modèles de régression logistique de départ :

- Durée du sinistre,
- Cause du sinistre,
- L'âge à la survenance du sinistre,
- L'âge au moment de l'effet du contrat,
- Le nombre de garanties souscrites,
- La durée entre la date d'effet du contrat et la survenance du sinistre,
- Le nombre de sinistres pour chaque contrat de la base (cette variable est créée en comptant le nombre de sinistres pour chaque contrat avant un nouvel arrêt, cela fait que pour un contrat avec plusieurs sinistres, nous n'aurons pas le même nombre de sinistres sur chaque ligne)
- Le sexe de l'assuré,
- Le top fumeur,
- Le niveau de sélection médicale,
- Le type de flux (il s'agit du canal de réception de la demande d'indemnisation dans CRM),
- Les franchises accident, maladie et hospitalisation,
- Le montant des indemnités journalières,
- Le réseau de distribution,
- Le code région portefeuille,
- La présence d'enfants,
- L'âge au moment de l'effet du contrat,
- Le produit.

Nous verrons dans la suite de notre étude leur impact sur la variable "OUTCOME" que nous souhaitons prédire.

Chapitre 5

Modélisation : prédiction de la décision pour un nouvel arrêt de travail

5.1 Division de notre base de données

Nous décidons de ne garder que les lignes qui étaient remplies pour éviter de créer un biais, ce qui fait que nous n'avons que des valeurs non renseignées pour toutes les variables liées aux garanties 5 et 6 (garanties complémentaires au RO 365 jours et 90 jours). Cela crée un biais, mais nous le considérons négligeable, car nous avons très peu de dossiers rattachés à ces garanties dans nos données.

Après cela, nous décidons de diviser notre base de données en 2 parties qui sont les suivantes :

- Une base d'apprentissage qui permettra aux différents modèles de s'entrainer et donc de vérifier leur qualité d'ajustement, nous décidons de dédier 60% de la base initiale (soit 19 800 lignes, 4000 refusés (donc 20% de dossiers refusés) et 15 800 acceptés (donc 80% de dossiers acceptés)) à cette partie,
- Une base de test qui permettra de vérifier le pouvoir explicatif de nos modèles, elle représentera aussi 40% de la base de départ (soit 13 200 lignes, 2 700 refusés (Soit 20,5% de dossiers refusés) et 10 500 acceptés (soit 79,5% de dossiers acceptés)).

5.2 Régression logistique

5.2.1 Premier modèle de régression logistique : le modèle complet

Étant donné la nature binaire de notre variable réponse, nous décidons de mettre en place un GLM binomial pour pouvoir mener à bien nos prédictions.
Les paramètres à prendre en compte pour ce modèle sont :

— La variable réponse : dans notre cas "OUTCOME" qui vaut 1 si le dossier est accepté et 0 sinon,

— Les variables explicatives : pour ce premier modèle, ce sont toutes les variables que nous avons décidé de garder à la suite du chapitre précédent sur l'étude de corrélation des variables.

- Durée du sinistre,
- Cause du sinistre,
- L'âge à la survenance du sinistre,
- L'âge au moment de l'effet du contrat,
- ...

  On pourra retrouver la liste complète dans la troisième partie du chapitre 4.

— Et la loi de la variable à prédire, ici Binomial().

Pour rappel, l'utilisation de variables qualitatives dans un modèle GLM nécessite la création de variables binaires pour chacune des modalités de la variable

en question.

Nous obtenons un AIC de 17 229,28 et observons les caractéristiques suivantes pour ce premier modèle :

	Echantillon d'apprentissage	Echantillon de test
VN	2 560	1 701
VP	10 902	7 245
FN	4 898	3 255
FP	1 440	999
VN+FP	4 000	2 700
FN+VP	15 800	10 500
Spécificité	0,64	0,63
Sensibilité	0,69	0,69
AUC	0,607	0,604

TABLE 5.1 – Caractéristique de la première régression logistique sur les données d'apprentissage et de test

Les résultats obtenus nous font penser que dans le cas où nous avons un dossier refusé et un dossier accepté, il y a 60% de chances que le modèle arrive à faire la distinction entre ces deux dossiers.

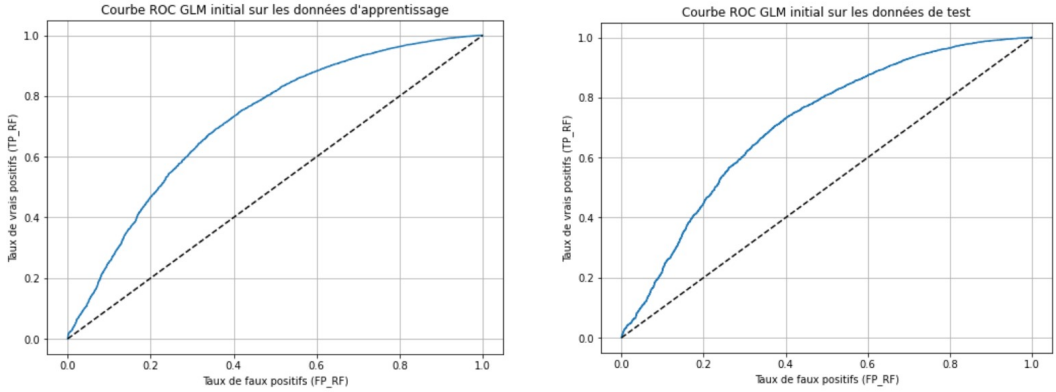


FIGURE 5.1 – Courbes ROC pour les données d'apprentissage et de test

5.2.2 Sélection des variables significatives et obtention du modèle de régression logistique optimisé

Afin d'améliorer les performances du GLM, nous utilisons les méthodes de sélection supervisée de variables *stepwise* et *backward*. Avec ces deux méthodes, nous obtenons que les variables qui expliquent le mieux le modèle (ainsi que leurs p-values),

nous permettant ainsi d'obtenir le modèle de régression logistique le plus performant sont les suivantes :

Variables	P_values
CAUSE_SINISTRE_VF_maladie	1.660074885032902e-89
DUREE_SINISTRE_60_	3.23862680740399e-87
DUREE_SINISTRE_30_45	8.798015999787167e-54
DUREE_SINISTRE_45_60	4.397714015133365e-52
DUREE_SINISTRE__15	1.9694817140523323e-32
DUREE_EFFET_AI__90	3.200344734117834e-29
FR_ACC_VF_365	1.2571633933586366e-11
CAUSE_SINISTRE_VF_accident_hospi	7.963935628727528e-9
FR_MAL_VF_30.0	1.3797424477031746e-8
DUREE_EFFET_AI_90_365	1.0034101687325353e-7
TYPE_FLUX_VF_Hors_Espace_Client	0.000001399396011906584
NBRE_GARANTIES	0.00000697842242521430 5
MONTANT_IJ_VF__100	0.000007688766232553626
FR_HOSPI_VF_30	0.00039534975506626415
CODE_REGION_PORTEFEUILLE_VF_SUD_OUEST	0.0006072382023736363
FR_HOSPI_VF_3	0.0007329202775638632
AGE_CONTRAT__33	0.0012205224537168522
FR_HOSPI_VF_60	0.004122188186151262
FR_MAL_VF_60.0	0.004122188186151628
FR_ACC_VF_60	0.004122188186151729
CODE_REGION_PORTEFEUILLE_VF_NORD_EST	0.004324022128319971
FR_ACC_VF_15	0.006832462691780388
FR_HOSPI_VF_15	0.006832462691782018
CODE_REGION_PORTEFEUILLE_VF_SUD_EST	0.0075554615547571095
GARANTIE_Garantie2	0.02689398064591462
NSM_NIVEAU2	0.028048036142990354
PRODUIT_PRODUIT2	0.03642020902802142
FR_ACC_VF_90	0.04909373177679665
FR_HOSPI_VF_90	0.049093731776798766
FR_MAL_VF_90.0	0.0490937317768016

TABLE 5.2 – P-values pour les variables les plus importantes du GLM

Nous constatons que toutes les variables explicatives citées ci-dessus sont statistiquement significatives et le modèle l'est aussi.

Nous pourrions donc dire qu'elles influencent le fait qu'après traitement, une demande d'indemnisation en arrêt de travail soit oui ou non acceptée.

En plus, les performances de ce modèle sont meilleures que celles de la première régression logistique, nous avons un AIC de 16 900 puis :

	Échantillon d'apprentissage	Échantillon de test
VN	2 556	1 701
VP	10 981	7 308
FN	4 819	3 192
FP	1 444	999
VN+FP	4 000	2 700
FN+VP	15 800	10 500
Spécificité	0,639	0,63
Sensibilité	0,695	0,696
AUC	0,627	0,623

TABLE 5.3 – Caractéristiques de la deuxième régression logistique sur les données d'apprentissage et de test

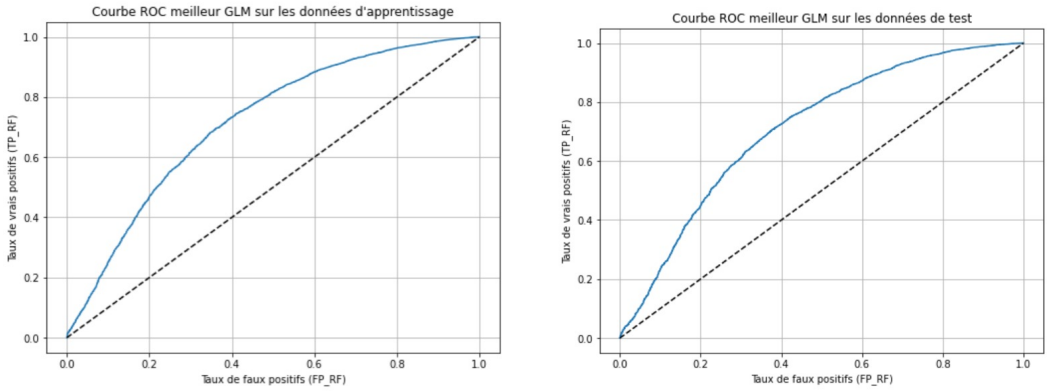


FIGURE 5.2 – Courbes ROC pour les données d'apprentissage et de test

Notre top 5 des variables explicatives les plus importantes pour ce modèle est le suivant :

1. Cause sinistre maladie,
2. Durée sinistre supérieure à 60 jours,
3. Durée du sinistre comprise entre 30 et 45 jours,
4. Durée du sinistre comprise entre 45 et 60 jours et
5. Durée du sinistre inférieure à 15 jours.

Ces résultats nous semblent cohérents, car intuitivement, nous pouvons nous attendre à ce que les dates de début et de fin d'un nouvel arrêt de travail soient très importantes dans la prise de décision des gestionnaires. En effet, ces informations peuvent

être décisives, car nous pouvons par exemple comparer la durée du sinistre aux potentielles franchises qui seraient rattachées au contrat de l'assuré et donc décider si finalement, nous acceptons ou pas d'indemniser son arrêt.

La cause du sinistre peut aussi s'avérer très importante, car c'est l'information qui nous permet de savoir quelle franchise appliquer lorsqu'on reçoit une nouvelle demande d'indemnisation.

Après cette étape, nous décidons de traiter le déséquilibre de la base de données pour voir si cela permet d'améliorer les performances du modèle.

5.2.3 Suréchantillonage de la classe minoritaire (dossiers refusés)

Pour cela, nous utilisons la technique *SMOTE* qui permet de traiter les déséquilibres de classe en suréchantillonant la classe minoritaire de nos données. Nous l'utilisons sur les données d'apprentissage, nous avions au départ 80% de dossiers acceptés (15 800 dossiers) et 20% de dossiers refusés (4 000 dossiers) et nous avons donc maintenant, 31 600 lignes dans notre base dont 15 800 concernent des dossiers acceptés (soit 50%) et 15 800 concernent des dossiers refusés (soit 50%).

	Echantillon d'apprentissage	Echantillon de test
VN	11 060	1 890
VP	10 428	7 035
FN	5 372	3 465
FP	4 740	810
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Spécificité	0,66	0,67
Sensibilité	0,7	0,7
AUC	0,64	0,64

TABLE 5.4 – Caractéristique de la deuxième régression logistique sur les données d'apprentissage et de test

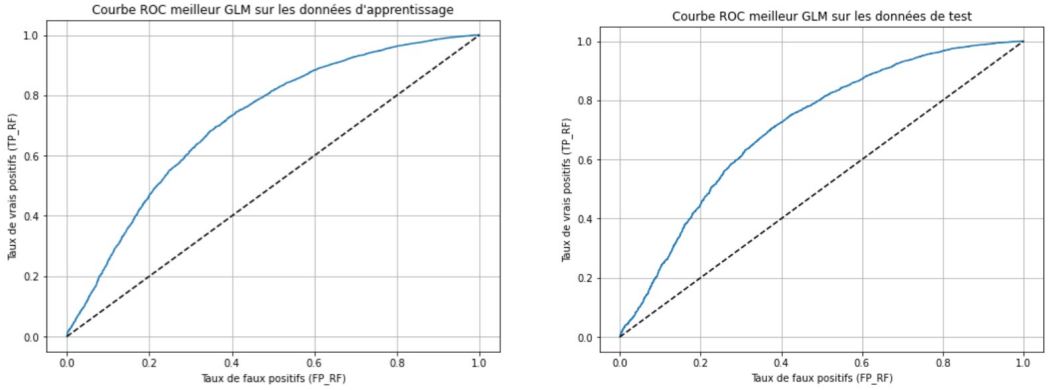


FIGURE 5.3 – Courbes ROC pour les données d'apprentissage suréchantillonnées et les données de test

Nous remarquons que le suréchantillonnage de la classe minoritaire nous permet d'augmenter légèrement l'AUC sur les données d'apprentissage et de test, mais également augmenter la sensibilité et la spécificité. Nous pourrons donc dire que cela nous permet d'améliorer le pouvoir prédictif de notre modèle, car celui-ci prédit un peu mieux les dossiers refusés maintenant que nos données sont équilibrées.

5.2.4 Prise en compte de l'interaction entre certaines variables

Nous décidons dans cette partie d'observer l'effet de la prise en compte de l'interaction entre certaines variables sur la qualité de notre modèle.
Les résultats de l'analyse descriptive bivariée nous poussent à nous intéresser aux interactions entre les variables ci-dessous :

— Cause du sinistre et Sexe de l'assuré et

— Âge à la survenance du sinistre et Sexe de l'assuré

Cependant, l'ajout de ces nouvelles variables explicatives à notre modèle n'apportant pas plus de précision et dégradant les performances de notre modèle, nous décidons de ne pas pousser l'étude plus loin.

5.2.5 Interprétation des résultats

Les résultats obtenus nous paraissent pertinents. En effet, nous obtenons des résultats auxquels nous nous attendions plus ou moins intuitivement. Ils nous emmènent à nous dire que les questions les plus importantes que doivent se poser les gestionnaires pour tout nouveau dossier d'arrêt de travail initial afin d'avoir une idée de la décision sont les suivantes :

— Est-ce la durée du sinistre est plutôt longue (plus de 60 jours)

— Ou est-ce qu'elle est moyennement longue (entre 30 et 45 jours ou entre 45 et 60 jours) ou est-ce qu'elle est très courte (moins de 15 jours) ?

Les odds ratios pour les variables les plus importantes

Les odds ratios sont des mesures statistiques qui permettent de quantifier l'association entre deux variables binaires et d'évaluer la force de cette association.

Pour une variable qualitative, ils s'interprètent comme suit :
On compare les modalités spécifiées dans le tableau résumé de notre modèle à la modalité de référence (il s'agit de la modalité de notre variable qui n'est pas renseignée dans le tableau résumé) et si l'odds ratio est supérieur à 1, cela veut dire que l'acceptation est plus fréquente chez la modalité spécifiée par rapport à la modalité de référence et sinon, cela veut dire le contraire.

Pour une variable quantitative, l'interprétation se fait selon une unité de différence, cela veut dire que pour le nombre de garanties souscrites par exemple, si l'odds ratio est supérieur à 1, l'évènement ici l'acceptation est plus fréquente quand on augmente le nombre de garanties et donc par extension, l'acceptation est plus fréquente pour les dossiers concernant des contrats ayant le plus grand nombre de garanties.

On trouvera ci-dessous un tableau représentant les modalités de référence pour chaque variable qualitative importante de notre modèle :

Variables	**Modalité de référence**
CAUSE_SINISTRE	accident
DUREE_SINISTRE	durée entre 15 et 30 jours
DUREE_EFFET_AI	plus de 365 jours
AGE_CONTRAT	entre 33 et 61 ans
NSM	niveau 1
MONTANT_IJ	entre 100 et 300
TYPE_FLUX	espace client
CODE_REGION_PORTEFEUILLE	DOMTOM
FR_ACC	3 jours
FR_MAL	7 jours
FR_HOSPI	0 jours
GARANTIE	garantie 1
PRODUIT	PRODUIT 1

TABLE 5.5 – Modalité de référence pour chaque variable explicative

Ils sont donc intéressants pour notre étude, car ils permettent d'établir une sorte de profil des dossiers qui sont davantage susceptibles d'être acceptés ou refusés.

Nous obtenons les odds ratios ci-dessous pour nos 5 variables plus importantes :

Variables	Odds Ratios
CAUSE_SINISTRE_VF_maladie	0.43036697731241996
DUREE_SINISTRE_60_	3.993955686714954
DUREE_SINISTRE_30_45	2.2799874163717235
DUREE_SINISTRE_45_60	3.265981012205225
DUREE_SINISTRE__15	0.547546180894591

TABLE 5.6 – Odds Ratio des 5 variables les plus importantes de la régression logistique

Nous pouvons interpréter ces odds ratios de la manière suivante :

— Il serait plus probable d'accepter un dossier quand la durée de celui-ci est strictement supérieure à 60 jours que lorsqu'elle est comprise entre 15 et 30 jours. Cela serait dû au fait que l'odds ratio des durées de sinistres de plus de 60 jours est égal à environ 4 fois celui des sinistres de durée comprise entre 15 et 30 jours.
Ce résultat nous paraît cohérent, car nous avions remarqué avec nos analyses que nous avions beaucoup moins de refus pour les sinistres de durées les plus longues. De plus, le fait que plus la durée du sinistre est élevée et moins, il est probable d'être refusé pour cause de franchise atteinte (le motif de refus le plus courant) nous conforte pour le résultat observé.

— De la même manière, nous pourrions aussi dire que nous aurions moins de chance d'accepter un dossier de durée inférieure ou égale à 15 jours que d'en accepter un de durée comprise entre 15 et 30 jours, ce qui encore une fois vient confirmer les résultats auquel nous nous attendions au vu de nos analyses.

Dans la suite, nous nous intéressons aux autres modèles de machine learning cités dans le chapitre 2 afin de comparer les résultats de notre GLM aux résultats de ces derniers.

5.3 Forêt aléatoire

Un modèle de forêt aléatoire sera mis en place dans cette partie pour la prédiction de la variable "OUTCOME" ; il contiendra toutes les variables utilisées pour le premier GLM. Et dans la suite, nous verrons comment se déterminent les coefficients d'importance des variables pour ce modèle et comment les interpréter.

5.3.1 Application d'un modèle de forêt aléatoire à nos données

Nous appliquons le modèle de forêt aléatoire à nos données. Une optimisation des hyper-paramètres est envisagée avec les fonctions *GridSearchCV*, *RandomizedSearchCV* et *Optuna*.

Les deux premières méthodes tournent sans s'arrêter et l'utilisation de Optuna donne des résultats mais ces derniers ne sont pas concluants.

Nous choisissons donc de procéder en tâtonnant et nous obtenons que les hyperparamètres les plus adaptés sont :

— nombre d'arbres = 1 000,

— profondeur = 30 et,

— nombre d'éléments minimum dans chaque nœud = 2.

Nous obtenons les caractéristiques ci-dessous pour le modèle complet (celui avec toutes les variables du premier GLM).

	Échantillon d'apprentissage	Échantillon de test
VN	3 600	1 134
VP	14 694	9 345
FN	1 106	1 155
FP	400	1 566
VN+FP	4 000	2 700
FN+VP	15 800	10 500
Spécificité	0,90	0,42
Sensibilité	0,93	0,89
AUC	0,91	0,68

TABLE 5.7 – Caractéristiques du modèle de forêt aléatoire sur les données d'apprentissage et de test

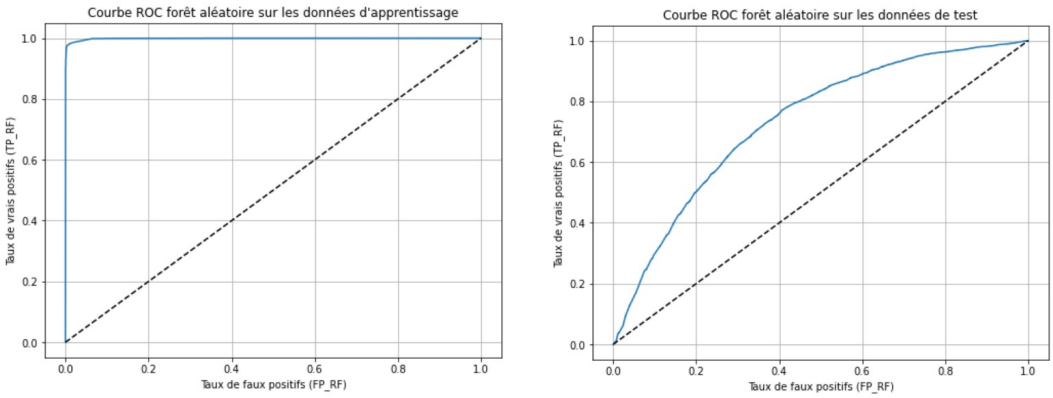


FIGURE 5.4 – Courbes ROC pour les données d'apprentissage et de test

La spécificité observée sur les données de test nous montre que le modèle de forêt aléatoire a du mal à prédire les dossiers refusés, mais il prédit bien les dossiers acceptés. En effet, sur 100 dossiers refusés, il arrive à en prédire seulement 42 ce qui

peut être grave, car prendre en charge des dossiers refusés à tort pourrait entrainer une hausse du S/C.

Nous décidons donc dans la suite d'observer les effets du suréchantillonnage de la classe minoritaire dans la base d'apprentissage sur nos prédictions.

5.3.2 performances du modèle sur les données après suréchantillonnage de la classe minoritaire

Après cette étape, nous décidons de traiter le déséquilibre de la base de données pour voir si cela permet d'améliorer les performances du modèle.

Nous utilisons les mêmes données que celles citées dans la partie 3 de la régression logistique et obtenons les résultats suivants :

	Échantillon d'apprentissage	Échantillon de test
VN	15 642	2 160
VP	15 484	8 190
FN	316	2 310
FP	158	540
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Spécificité	0,98	0,78
Sensibilité	0,99	0,8
AUC	0,99	0,89

TABLE 5.8 – Caractéristique du modèle de forêt aléatoire sur les données d'apprentissage et de test

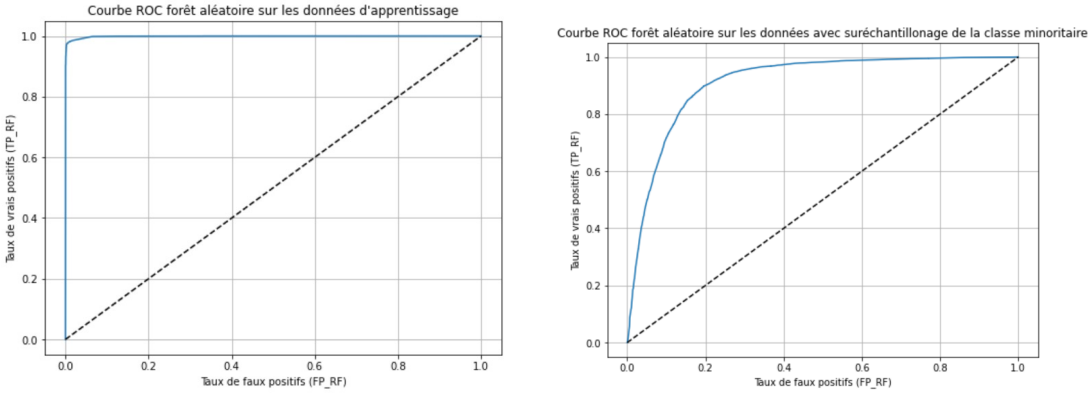


FIGURE 5.5 – Courbes ROC pour les données d'apprentissage après suréchantillonnage et les données de test

Nous pouvons dire que cela nous permet d'améliorer les performances de notre modèle, en effet, nous augmentons sa capacité à prédire les dossiers refusés sur la base de test, sur 100 dossiers refusés, nous pouvons prédire 78 correctement alors qu'auparavant, nous n'arrivions à en prédire que 42 correctement.

Nous déterminons maintenant les scores d'importance des variables de notre modèle. Plus les scores d'importance sont élevés et plus les variables sont importantes. Nous obtenons les scores d'importance suivants (nous avons conservé les variables avec des scores supérieurs à 0) :

Variables	Score d'importance
CAUSE_SINISTRE_VF_maladie	$0.032+/-0.003$
DUREE_SINISTRE_60_	$0.026+/-0.001$
DUREE_SINISTRE__15	$0.020+/-0.002$
DUREE_SINISTRE_30_45	$0.018+/-0.002$
DUREE_SINISTRE_45_60	$0.011+/-0.001$
NBRE_GARANTIES	$0.005+/-0.001$
DUREE_EFFET_AI__90	$0.004+/-0.001$
SEXE_VF_Masculin	$0.004+/-0.001$
TOP_FUMEUR_non	$0.003+/-0.001$
AGE_CONTRAT__33	$0.003+/-0.001$
PRESENCE_ENFANT_VF_OUI	$0.003+/-0.001$
RESEAU_VF_RESEAU2	$0.002+/-0.001$
CLAUSE_VF_OUI	$0.002+/-0.001$
CAUSE_SINISTRE_VF_maladie_hospi	$0.002+/-0.001$
NSM_NIVEAU3	$0.002+/-0.001$
DUREE_EFFET_AI_90_365	$0.002+/-0.001$
MONTANT_IJ_VF__100	$0.001+/-0.001$
TYPE_FLUX_Hors_Espace_Client	$0.001+/-0.001$
CODE_REGION_PORTEFEUILLE_VF_SUD_OUEST	$0.001+/-0.001$
CAUSE_SINISTRE_VF_accident_hospi	$0.001+/-0.000$
CODE_REGION_PORTEFEUILLE_VF_ILE_DE_FRANCE	$0.001+/-0.000$
FR_HOSPI_VF_30	$0.001+/-0.000$

TABLE 5.9 – Scores d'importance

Dans le top 5 des variables les plus importantes pour notre modèle de forêt aléatoire, on retrouve exactement les mêmes variables que pour le meilleur GLM. Néanmoins, nous remarquons un décalage entre la place de la variable "durée du sinistre inférieure ou égale à 15 jours" dans le top 5 des variables les plus importantes pour la forêt aléatoire et sa place dans le top 5 pour le GLM. En effet, cette variable se trouve respectivement en cinquième et troisième place dans le top 5 des variables les plus importantes du Random Forest et du GLM.

Nous observons une dégradation des performances de notre modèle quand on essaie de sélectionner les variables explicatives par rapport à leurs coefficients d'importance. Nous décidons donc de garder toutes les variables explicatives dans notre modèle.

5.3.3 Analyse et interprétation des résultats

Nous observons des résultats d'importance de variables quelque peu différents de ceux obtenus avec le GLM binomial. Cependant, il est aussi important de noter que sur les données d'apprentissage de départ, le modèle de forêt aléatoire est très efficace alors que sur

les données de test de départ, il l'est beaucoup moins, mais cela est dû de toute évidence au déséquilibre de la base, car, sur les données après suréchantillonnage, nous constatons une nette amélioration de la capacité du modèle à prédire les dossiers refusés.

5.4 Réseau de neurones

5.4.1 Application du réseau de neurones à nos données

Nous appliquons le modèle de réseau de neurones pour les variables réponses de type **catégorielles** à nos données. Comme pour le modèle de forêt aléatoire, on envisage les deux premières techniques citées pour l'amélioration des hyperparamètres, mais encore une fois, le temps d'exécution est trop élevé. Nous décidons alors de tester Optuna, cependant, cela ne donnant pas de résultats optimaux, nous choisissons de tâtonner pour trouver des hyperparamètres adaptés.
Les hyperparamètres que nous obtenons sont les suivants :

- Nombre de couches cachées : 3,
- Nombre de neurones dans chaque couche cachée : 30, 60 et 50,
- Fonction d'activation : 'logistic',
- Méthode d'optimisation des poids : 'adam' et,
- Maximum d'itérations pour la convergence du modèle : 200.

Une fois le modèle mis en place, nous obtenons les statistiques suivantes :

	Échantillon d'apprentissage	Échantillon de test
VN	2 400	864
VP	14 852	9 135
FN	948	1 365
FP	1 600	1 836
VN+FP	4 000	2 700
FN+VP	15 800	10 500
Spécificité	0,6	0,32
Ssensibilité	0,94	0,87
AUC	0,80	0,6228

TABLE 5.10 – Caractéristiques du réseau de neurones sur les données d'apprentissage et de test

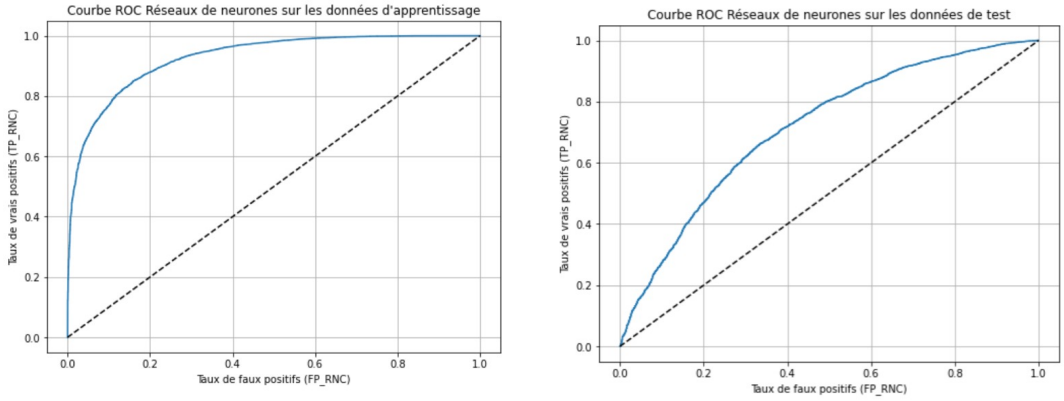


FIGURE 5.6 – Courbes ROC pour les données d'apprentissage et de test

Ces différents éléments nous montrent que le modèle de réseau de neurones n'est pas du tout efficace pour prédire les dossiers refusés. En effet, sur 100 dossiers refusés, il arrive à en prédire seulement 32 correctement, cela n'est pas du tout satisfaisant.

Nous décidons donc, comme nous l'avions fait pour le modèle de forêt aléatoire, de voir ce que l'on obtient avec les données d'apprentissage après suréchantillonnage.

5.4.2 Performances du modèle après suréchantillonnage de la classe minoritaire

Nous utilisons ici la base d'apprentissage suréchantillonnée et nous observons les résultats suivants :

	Echantillon d'apprentissage	Échantillon de test
VN	14 536	2 187
VP	14 062	7 665
FN	1 738	2 835
FP	1 264	513
VN+FP	15 800	2 700
FN+VP	15 800	10 500
Spécificité	0,89	0,73
Sensibilité	0,92	0,81
AUC	0,91	0,82

TABLE 5.11 – Caractéristiques du réseau de neurones sur les données d'apprentissage après suréchantillonage et les données de test

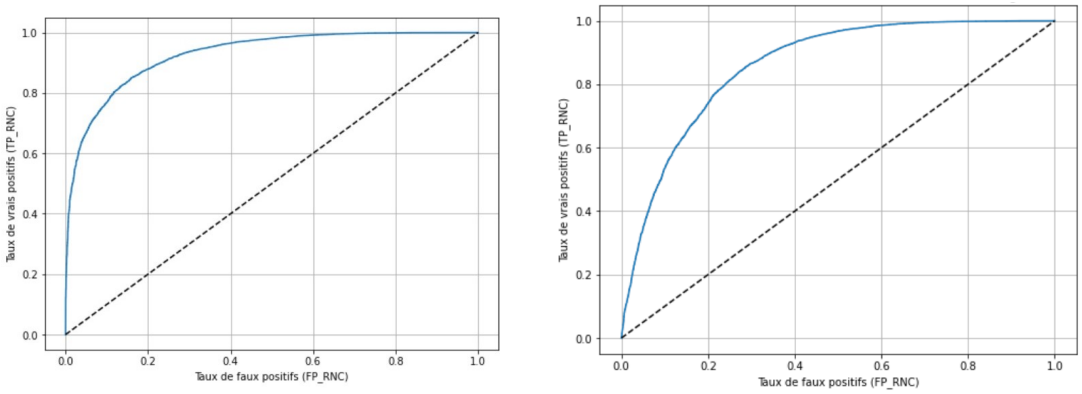


FIGURE 5.7 – Courbes ROC pour les données d'apprentissage après suréchantillonnage et les données de test

Nous constatons que rééquilibrer la base d'apprentissage nous permet d'améliorer la qualité des prédictions du modèle, nous arrivons à mieux prédire les dossiers refusés.

Pour évaluer l'importance des variables dans le réseau de neurones, nous utilisons la méthode de permutation qui consiste à mélanger les valeurs d'une variable à la fois dans les données test et à observer comment cela affecte les performances du modèle. Nous n'obtenons pas les mêmes variables les plus importantes pour ce modèle que celles obtenues avec le modèle de forêt aléatoire qu'on retrouve dans le tableau 5.9. Le top 5 des variables les plus importantes est le suivant :

Variable	Importance
NBRE_GARANTIES	0.08068951402842892
CAUSE_SINISTRE_VF_accident_hospi	0.054774131130460604
CAUSE_SINISTRE_VF_maladie	0.05293592319714218
CAUSE_SINISTRE_VF_maladie_hospi	0.04598124581379769
MONTANT_IJ_VF_300_	.04068244399791

TABLE 5.12 – Tableau du top 5 des variables les plus importantes selon le réseau de neurones

5.4.3 Analyse des résultats

Les résultats obtenus avec ce modèle (que nous retrouvons dans l'annexe D) sont différents de ceux que nous obtenons avec la forêt aléatoire et le GLM, pour trancher, nous allons donc réaliser une comparaison des performances de ces 3 modèles afin de pouvoir trancher et identifier celui qui a les meilleures performances.

	GLM	**Forêt aléatoire**	**Réseau de neurones**
AUC	0,64	0,89	0,82
Sensibilité	0,7	0,8	0,81
Spécificité	0,67	0,78	0,73

TABLE 5.13 – Récapitulatif des résultats de nos modèles sur la base de test

5.5 Comparaison des différents modèles et choix du plus adapté

En observant les sensibilités, spécificités et AUC de nos différents modèles après entrainement avec les données d'apprentissage suréchantillonnées (confère tableau 5.13), nous pouvons supposer que le modèle le plus adapté à notre étude est la forêt aléatoire. Avant, avec les données déséquilibrées, nous aurions plutôt choisi le modèle de régression logistique comme meilleur modèle vu qu'il avait la meilleure spécificité sur les données de test.
Au vu des résultats observés, nous pourrions choisir la forêt aléatoire, car par rapport à l'aspect actuariel même, c'est avec ce modèle qu'il serait moins risqué de prédire les dossiers refusés. Choisir le GLM ne serait pas une décision intéressante, car c'est le modèle avec lequel nous aurions un plus grand risque de prédire des dossiers comme acceptés alors qu'ils devraient être refusés. Procéder de cette manière ferait augmenter les charges de l'assureur alors que son objectif est de les réduire.
Cependant, il est important de prendre en compte le fait que l'utilisation du suréchantillonnage de la classe minoritaire pour l'amélioration des performances prédictives de nos modèles peut créer un biais car les données ne reflètent plus vraiment la réalité.

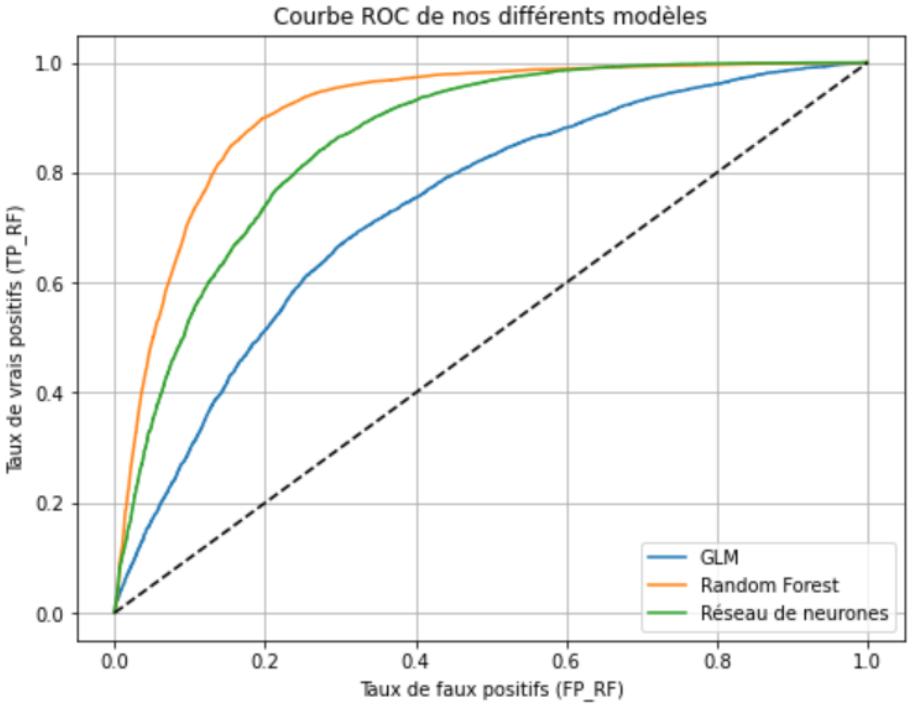


FIGURE 5.8 – Courbes ROC de nos 3 modèles sur les données de test

Les courbes ROC obtenues viennent confirmer nos précédentes interprétations.

Profil de nos dossiers

Il faut noter que le glm bien qu'étant un peu plus compliqué à mettre en place et ayant les performances de prédiction les moins bonnes, il représente un avantage considérable, car permettant d'obtenir de manière relativement facile les odds ratio des variables explicatives, ce qui nous aide à établir une liste des caractéristiques que nous pourrions retrouver le plus chez les dossiers refusés que chez les acceptés (et inversement). Le tableau complet des odds ratio des variables les plus importantes pour le GLM (Annexe C) nous emmène à établir le profil suivant selon les caractéristiques d'un nouvel arrêt de travail :

— Plus le nombre de garanties souscrites au moment du sinistre est élevé et plus, il y aurait de chance que le dossier soit accepté, nous avons 1,12 fois plus de chance d'accepter un dossier avec 2 garanties qu'un dossier avec une seule garantie,

— Si le dossier concerne un arrêt dû à un accident nécessitant une hospitalisation, alors il aurait près de 3 fois plus de chance d'être accepté que si l'arrêt est dû à la cause accident simple et s'il est dû à une maladie, alors il aurait seulement 0,43 fois plus de chance d'être accepté que s'il est dû à un accident, cela voudrait dire qu'il aurait moins de chance d'être accepté par rapport au sinistre causé par l'accident,

— Si le montant des indemnités journalières garanti est inférieur à 100 €, alors l'acceptation du dossier serait moins fréquente que lorsqu'il est compris entre 100 et 300 €,

— Les dossiers concernant les franchises accident de 15 jours seraient davantage refusés que ceux qui concernent la franchise accident 3 jours,

— Si pour un dossier donné, la franchise maladie est de 30 jours, alors il aurait 0,70 fois plus de chance d'être accepté que si la franchise maladie est de 7 jours, cela voudrait dire qu'il serait plus fréquent de refuser un sinistre si la franchise maladie est de 30 jours que si elle est de 7 jours,

— Un dossier pour lequel la franchise hospitalisation est de 3 jours, serait plus susceptible d'être refusé que celui qui a une franchise hospitalisation de 0,

— Si le niveau de sélection médicale rattaché au contrat est le niveau 2 alors, le dossier serait plus susceptible d'être accepté que s'il s'agit du niveau 1,

— Les dossiers rattachés aux contrats qui concernent le produit 2 seraient plus fréquemment acceptés que ceux qui concernent les dossiers rattachés au produit 1. Il se pourrait cependant que cela soit lié au volume de dossiers rattachés au produit 1. En effet, il s'agit du produit le plus présent dans notre base de données,

— Les dossiers pour lesquels la durée entre la date d'effet du contrat et la date de survenance du sinistre est inférieure au délai d'attente (90 ou 365 jours en fonction du produit) seraient davantage susceptibles d'être refusés que ceux pour lesquels elle est supérieure à 365 jours. Cela est normal vu que c'est l'utilité même du délai d'attente.

Conclusion

Bilan

Au terme de cette étude, nous pouvons dire dans un premier temps que les différentes méthodes que nous avons appliquées nous ont permis d'avoir des résultats qui nous semblent plutôt cohérents. En effet, nous avions intuitivement pensé à certains résultats avant même d'entamer l'étude et ces derniers ont été confirmés.

Dans un second temps, nous avons pu améliorer les performances prédictives de nos modèles et donc trouver ceux qui seraient plus adaptés pour la prédiction des deux classes de notre OUTCOME.
Même si nous n'avons pas été à mesure d'améliorer de manière plus efficace les performances de la régression logistique, nous pouvons dire que les résultats obtenus nous donnent une piste sur le profil de nos dossiers acceptés (et inversement refusés) et les éléments de ce profil pourraient aider les gestionnaires à se poser des questions plus pertinentes qui pourraient faciliter leurs analyses et donc leur prise de décision sur des nouvelles demandes d'ouverture en arrêt de travail.

Cependant, il serait précipité d'affirmer que ces caractéristiques que nous avons déterminées sont exhaustives, car il y a des variables qui nous semblaient très intéressants, mais que nous n'avons pas été à mesure d'exploiter, telles que les professions et les codes CSP pour lesquels nous avons été confrontés à différentes difficultés que nous aborderons dans la partie suivante.

Difficultés rencontrées, limites et action à entreprendre en vue d'une amélioration

Nous avons rencontré plusieurs difficultés pour la création des nouvelles variables. En effet, cette étape de l'étude a été la plus compliquée, car les commentaires CRM n'étaient pas structurés et qu'ils ne contenaient pas les informations (nous avions des formats de dates très compliqués à gérer et parfois des commentaires trop courts pour être exploités, surtout pour les dossiers refusés par exemple) dont nous avions besoin dans certains cas.

Nous avons également rencontré plusieurs difficultés qui nous ont poussés à renoncer à l'exploitation de certaines variables :

- Une quantité astronomique de modalités pour la profession et le code CSP (plus de 3 000 modalités pour ces variables) dans notre jeu de données : il aurait été intéressant d'observer les effets de la profession sur la décision d'accepter ou non un arrêt de travail, car certains métiers sont plus risqués que d'autres. En effet, on a tendance à penser qu'une personne travaillant dans le secteur du bâtiment a plus de chance d'être victime d'un accident qu'une personne travaillant dans le secteur de l'assurance par exemple, de plus, les hommes et les femmes n'exerçant pas forcément les mêmes professions, l'interaction entre le sexe de l'assuré et la profession aurait pu nous donner une information supplémentaire sur les caractéristiques principales des dossiers qui ont plus de chance d'être acceptés que refusés.
- Cependant, nous n'aurions toujours pas pu les intégrer à nos analyses, car la seule méthode envisageable aurait été de regrouper certaines modalités, mais cela serait compliqué, car le code CSP peut être renseigné différemment d'un produit à un autre.

Annexe

Annexe A

Étapes de la jointure des données CRM aux données VANDA

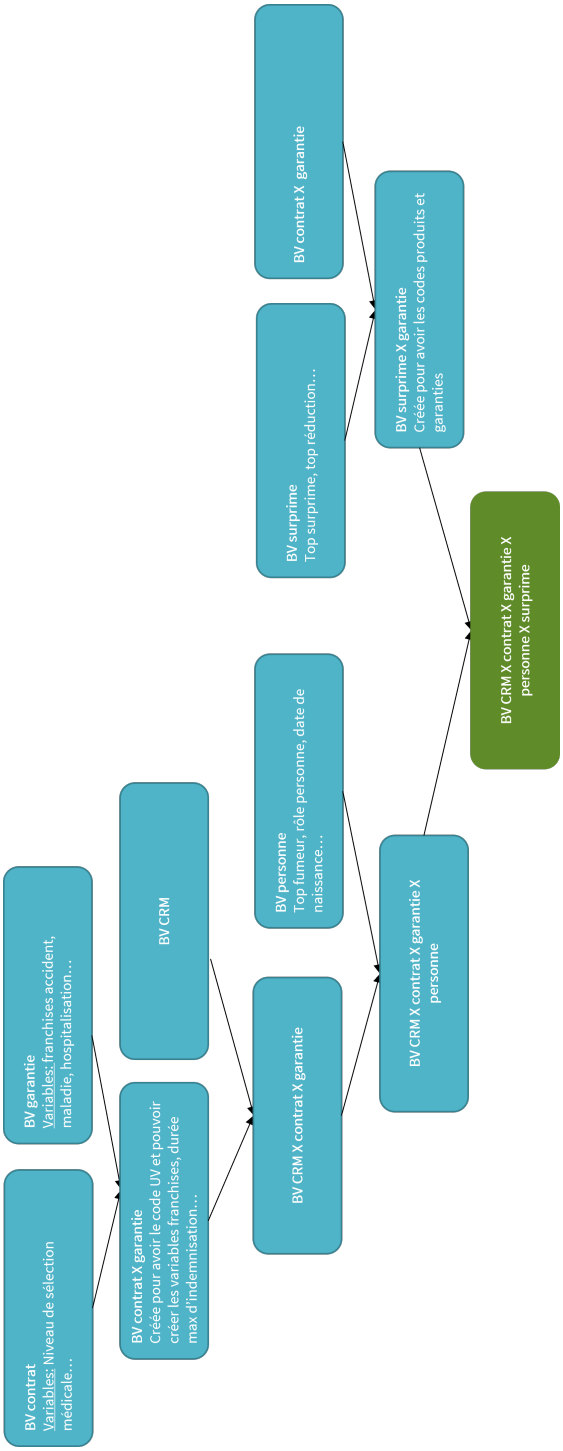


FIGURE A.1 – Première étape des jointures des données CRM aux données VANDA

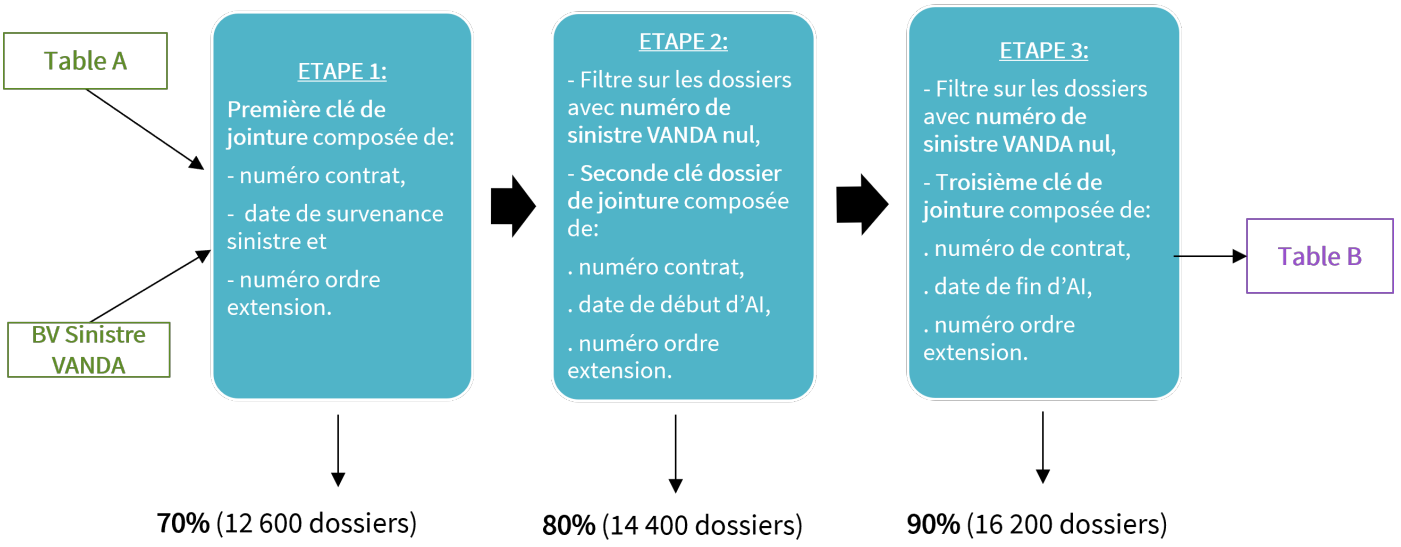


FIGURE A.2 – Deuxième étape des jointures des données CRM aux données VANDA

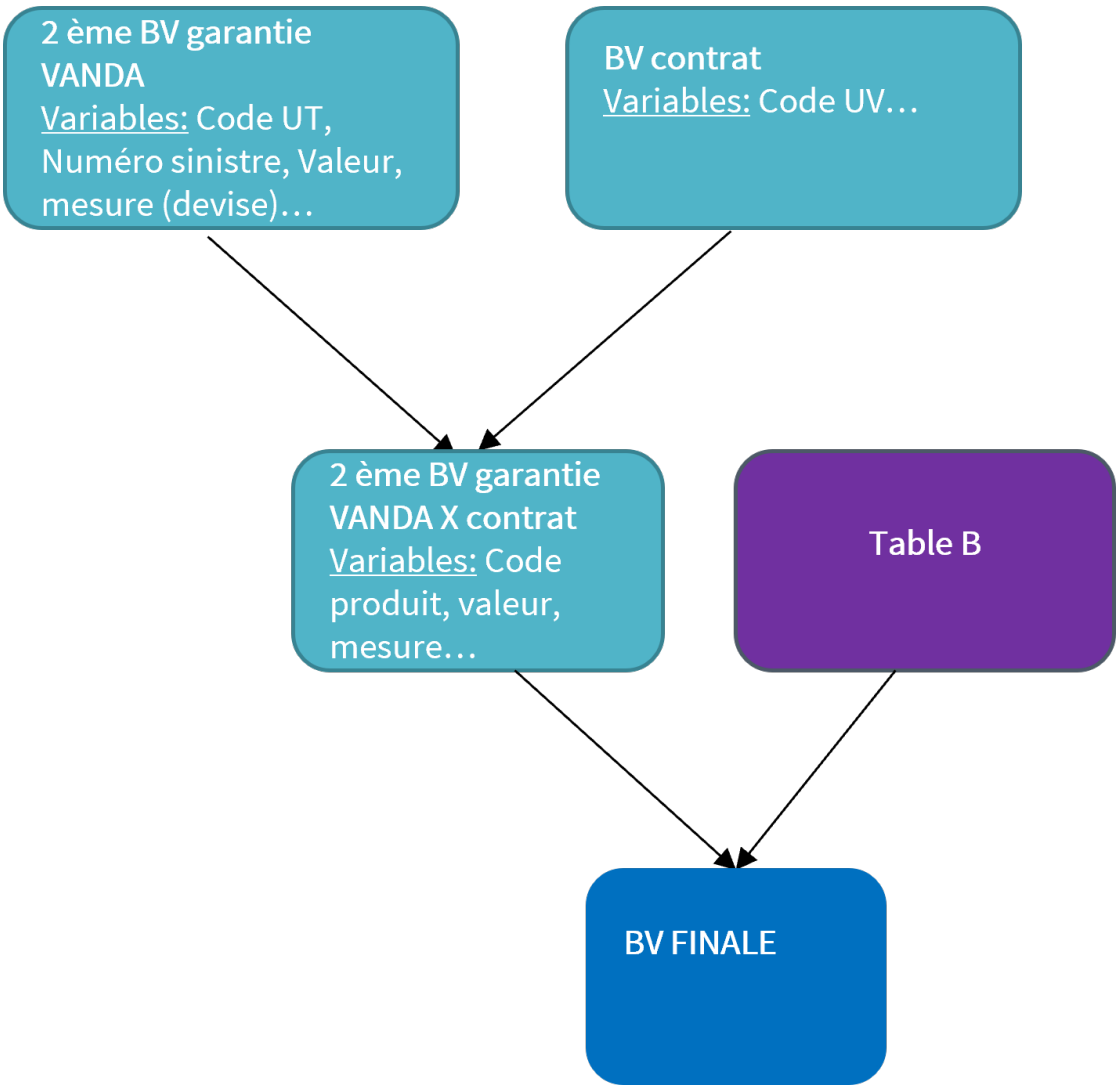


FIGURE A.3 – Troisième étape des jointures des données CRM aux données VANDA

Annexe B

Résultats complets du test de Khi 2

Variable 1	Variable 2	Chi-squared	P-value
MONTANT_IJ_VF	STATUT_CONTRAT_VF	4,85	0,77
FR_ACC_VF	TOP_REFUS_AV_DECLA	10,30	0,17
FR_MAL_VF	TOP_REFUS_AV_DECLA	9,94	0,27
FR_HOSPI_VF	TOP_REFUS_AV_DECLA	13,64	0,09
DUREE_MAX_INDM_VF	NBRE_SINISTRES_VF	0,52	0,47
DUREE_MAX_INDM_VF	TOP_REFUS_AV_DECLA	1,01	0,31
DUREE_MAX_INDM_VF	LIBELLE_ETAT_EXO_INCAP	10,76	0,15
DUREE_MAX_INDM_VF	SEXE_VF	0,00	0,98
DUREE_MAX_INDM_VF	NSM	6,75	0,15
DUREE_MAX_INDM_VF	TOP_FUMEUR_VF	0,17	0,68
DUREE_MAX_INDM_VF	RESEAU_VF	3,62	0,16
DUREE_MAX_INDM_VF	MOTIF_STATUT_VF	20,70	0,87
DUREE_MAX_INDM_VF	TYPE_FLUX_VF	0,61	0,43
DUREE_MAX_INDM_VF	STATUT_VF	3,56	0,06
DUREE_MAX_INDM_VF	PRESENCE_ENFANT_VF	1,70	0,19
DUREE_MAX_INDM_VF	CLAUSE_VF	0,83	0,36
DUREE_MAX_INDM_VF	AGE_SURVENANCE_SINISTRE	3,41	0,33
DUREE_MAX_INDM_VF	AGE_FIN_AI	3,36	0,34
DUREE_MAX_INDM_VF	AGE_CONTRAT	0,84	0,84
DUREE_MAX_INDM_VF	DUREE_EFFET_AI	4,83	0,09
LIBELLE_ETAT_GARANTIE	STATUT_VF	8,79	0,19
LIBELLE_ETAT_EXO_INCAP	DUREE_MAX_INDM_VF	10,76	0,15
STATUT_CONTRAT_VF	MONTANT_IJ_VF	4,85	0,77
STATUT_CONTRAT_VF	TOP_SURPRIME_VF	0,44	0,98
SEXE_VF	DUREE_MAX_INDM_VF	0,00	0,98
TOP_SURPRIME_VF	NBRE_SINISTRES_VF	0,01	0,91
TOP_SURPRIME_VF	TOP_REFUS_AV_DECLA	0,99	0,32
TOP_SURPRIME_VF	STATUT_CONTRAT_VF	0,44	0,98
TOP_SURPRIME_VF	STATUT_VF	1,69	0,19
TOP_SURPRIME_VF	DUREE_HOSPI	2,92	0,71
NSM	DUREE_MAX_INDM_VF	6,75	0,15
NSM	STATUT_VF	2,89	0,58
TOP_FUMEUR_VF	TOP_HOSPI_VF	1,83	0,18
TOP_FUMEUR_VF	NBRE_SINISTRES_VF	0,06	0,81
TOP_FUMEUR_VF	DUREE_MAX_INDM_VF	0,17	0,68
TOP_FUMEUR_VF	PRESENCE_ENFANT_VF	0,20	0,65
TOP_FUMEUR_VF	DUREE_HOSPI	10,62	0,06
TOP_FUMEUR_VF	PRODUIT	6,17	0,10

TABLE B.1 – Suite des résultats du test de Khi2 (1/3)

Variable 1	Variable 2	Chi-squared	P-value
RESEAU_VF	DUREE_MAX_INDM_VF	3,62	0,16
RESEAU_VF	DUREE_HOSPI	12,01	0,28
CODE_REGION_PORTEFEUILLE_VF	TOP_REFUS_AV_DECLA	6,42	0,27
MOTIF_STATUT_VF	DUREE_MAX_INDM_VF	20,70	0,87
MOTIF_STATUT_VF	AGE_CONTRAT	108,64	0,06
TYPE_FLUX_VF	DUREE_MAX_INDM_VF	0,61	0,43
TYPE_FLUX_VF	PRESENCE_ENFANT_VF	1,61	0,20
STATUT_VF	DUREE_MAX_INDM_VF	3,56	0,06
STATUT_VF	LIBELLE_ETAT_GARANTIE	8,79	0,19
STATUT_VF	TOP_SURPRIME_VF	1,69	0,19
STATUT_VF	NSM	2,89	0,58
STATUT_VF	PRESENCE_ENFANT_VF	0,87	0,35
STATUT_VF	AGE_CONTRAT	7,11	0,07
STATUT_VF	GARANTIE	3,50	0,48
PRESENCE_ENFANT_VF	NBRE_SINISTRES_VF	0,02	0,89
PRESENCE_ENFANT_VF	TOP_REFUS_AV_DECLA	2,45	0,12
PRESENCE_ENFANT_VF	DUREE_MAX_INDM_VF	1,70	0,19
PRESENCE_ENFANT_VF	TOP_FUMEUR_VF	0,20	0,65
PRESENCE_ENFANT_VF	TYPE_FLUX_VF	1,61	0,20
PRESENCE_ENFANT_VF	STATUT_VF	0,87	0,35
PRESENCE_ENFANT_VF	DUREE_HOSPI	5,08	0,41
PRESENCE_ENFANT_VF	GARANTIE	9,30	0,05
TOP_NOUVELLE_ADHESION	CAUSE_SINISTRE_VF	11,07	0,09
TOP_NOUVELLE_ADHESION	TOP_HOSPI_VF	1,51	0,47
TOP_NOUVELLE_ADHESION	TOP_REFUS_AV_DECLA	4,06	0,13
TOP_NOUVELLE_ADHESION	DUREE_HOSPI	8,46	0,58
TOP_NOUVELLE_ADHESION	AGE_CONTRAT	8,97	0,18
CLAUSE_VF	DUREE_MAX_INDM_VF	0,83	0,36
DUREE_HOSPI	TOP_REFUS_AV_DECLA	8,97	0,11
DUREE_HOSPI	TOP_SURPRIME_VF	2,92	0,71
DUREE_HOSPI	TOP_FUMEUR_VF	10,62	0,06
DUREE_HOSPI	RESEAU_VF	12,01	0,28
DUREE_HOSPI	PRESENCE_ENFANT_VF	5,08	0,41
DUREE_HOSPI	TOP_NOUVELLE_ADHESION	8,46	0,58

TABLE B.2 – Suite des résultats du test de Khi2 (2/3)

Variable 1	Variable 2	Chi-squared	P-value
AGE_SURVENANCE_SINISTRE	DUREE_MAX_INDM_VF	3,41	0,33
AGE_FIN_AI	DUREE_MAX_INDM_VF	3,36	0,34
AGE_CONTRAT	TOP_REFUS_AV_DECLA	3,23	0,36
AGE_CONTRAT	DUREE_MAX_INDM_VF	0,84	0,84
AGE_CONTRAT	MOTIF_STATUT_VF	108,64	0,06
AGE_CONTRAT	STATUT_VF	7,11	0,07
AGE_CONTRAT	TOP_NOUVELLE_ADHESION	8,97	0,18
AGE_CONTRAT	PRODUIT	12,79	0,17
PRODUIT	TOP_FUMEUR_VF	6,17	0,10
PRODUIT	AGE_CONTRAT	12,79	0,17
DUREE_EFFET_AI	DUREE_MAX_INDM_VF	4,83	0,09
GARANTIE	TOP_REFUS_AV_DECLA	1,92	0,75
GARANTIE	STATUT_VF	3,50	0,48
GARANTIE	PRESENCE_ENFANT_VF	9,30	0,05

TABLE B.3 – Suite des résultats du test de Khi2 (3/3)

Annexe C

Odds Ratios

Variables	Odds_ratio
NBRE_GARANTIES	1.1201900450611406
CAUSE_SINISTRE_VF_accident_hospi	2.9815787887020986
CAUSE_SINISTRE_VF_maladie	0.43036697731241996
MONTANT_IJ_VF__100	0.7611025077420193
FR_ACC_VF_15	0.8811015831774455
FR_ACC_VF_60	0.7819647267869027
FR_ACC_VF_90	0.9427539093884633
FR_ACC_VF_365	0.617777680414499
FR_MAL_VF_30.0	0.7079545369671477
FR_MAL_VF_60.0	0.7819647267869028
FR_MAL_VF_90.0	0.9427539093884646
FR_HOSPI_VF_3	0.8098734544356924
FR_HOSPI_VF_15	0.8811015831774478
FR_HOSPI_VF_30	0.7461367739556032
FR_HOSPI_VF_60	0.7819647267869014
FR_HOSPI_VF_90	0.942753909388464
NSM_NIVEAU2	1.1355237221964256
CODE_REGION_PORTEFEUILLE_VF_NORD_EST	1.17953532828606
CODE_REGION_PORTEFEUILLE_VF_SUD_EST	1.1489167861783123
CODE_REGION_PORTEFEUILLE_VF_SUD_OUEST	1.1901910206161264
TYPE_FLUX_VF_Hors_Espace_Client	0.8228158504913884
DUREE_SINISTRE_30_45	2.2799874163717235
DUREE_SINISTRE_45_60	3.265981012205225
DUREE_SINISTRE_60_	3.993955686714954
DUREE_SINISTRE__15	0.547546180894591
AGE_CONTRAT__33	1.146023637368872
PRODUIT_PRODUIT2	1.207403459663962
DUREE_EFFET_AI_90_365	0.7469522100791893
DUREE_EFFET_AI__90	0.34403381435999647
GARANTIE_Garantie2	0.8911139130622152

TABLE C.1 – Tableau complet des odds Ratio

Annexe D

Importance des variables avec le réseau de neurones

Variables	Importance
NBRE_GARANTIES	0.08068951402842892
CAUSE_SINISTRE_VF_accident_hospi	0.054774131130460604
CAUSE_SINISTRE_VF_maladie	0.05293592319714218
CAUSE_SINISTRE_VF_maladie_hospi	0.04598124581379769
MONTANT_IJ_VF_300_	0.04068244399791614
MONTANT_IJ_VF__100	0.0397131056039294
FR_ACC_VF_7	0.03958844980278329
FR_ACC_VF_15	0.03743022996204505
FR_ACC_VF_30	0.03737069286299019
FR_ACC_VF_60	0.036272977599166416
FR_ACC_VF_90	0.03299843715114975
FR_ACC_VF_180	0.03223003646647312
FR_ACC_VF_365	0.03162350226985186
FR_MAL_VF_15.0	0.03141512242315989
FR_MAL_VF_30.0	0.030726724715338195
FR_MAL_VF_33.0	0.029828086626479074
FR_MAL_VF_60.0	0.02956389074942317
FR_MAL_VF_90.0	0.029495050978641003
FR_MAL_VF_180.0	0.028802932202128394
FR_MAL_VF_365.0	0.028533154722036106
FR_MAL_VF_VI	0.027669866785740817
FR_HOSPI_VF_1	0.026268884423606413
FR_HOSPI_VF_3	0.02524931160229212
FR_HOSPI_VF_15	0.024771154275507867
FR_HOSPI_VF_30	0.024218575574905057
FR_HOSPI_VF_60	0.02190965245218423
FR_HOSPI_VF_90	0.020705886730668994
FR_HOSPI_VF_180	0.020209124060430086
FR_HOSPI_VF_365	0.01832068169978411
SEXE_VF_Masculin	0.018303936890674943

TABLE D.1 – Importance des variables pour le réseau de neurones

Liste des tableaux

1 Caractéristique de la deuxième régression logistique sur les données d'apprentissage et de test . . . vii
2 Caractéristique du modèle de forêt aléatoire sur les données d'apprentissage et de test . . . vii
3 Caractéristiques du réseau de neurones sur les données d'apprentissage après suréchantillonage et les données de test . . . viii

4 Characteristic of second logistic regression on training and test data . . . x
5 Characteristic of random forest model on training and test data . . . xi
6 Features of the neural network on training data after oversampling and test data . . . xi

3.1 Récapitulatif des variables de notre base de données (1/2) . . . 37
3.2 Récapitulatif des variables de notre base de données (2/2) . . . 38
3.3 Taux de correspondance entre CRM et VANDA(1/4) . . . 48
3.4 Taux de correspondance entre CRM et VANDA(2/4) . . . 49
3.5 Taux de correspondance entre CRM et VANDA(3/4) . . . 50
3.6 Taux de correspondance entre CRM et VANDA(4/4) . . . 51
3.7 Taux de complétude des variables finales (1/3) . . . 54
3.8 Taux de complétude des variables finales (2/3) . . . 55
3.9 Taux de complétude des variables finales (3/3) . . . 56
3.10 Complétion des valeurs manquantes (1/2) . . . 61
3.11 Complétion des valeurs manquantes (2/2) . . . 61

4.1 Tableau de corrélation des variables (1/3) . . . 82
4.2 Tableau de corrélation des variables (2/3) . . . 83
4.3 Tableau de corrélation des variables (3/3) . . . 84
4.4 Aperçu des résultats du test de Khi 2 . . . 86

5.1 Caractéristique de la première régression logistique sur les données d'apprentissage et de test . . . 91
5.2 P-values pour les variables les plus importantes du GLM . . . 92

5.3 Caractéristiques de la deuxième régression logistique sur les données d'apprentissage et de test . . . 93
5.4 Caractéristique de la deuxième régression logistique sur les données d'apprentissage et de test . . . 94
5.5 Modalité de référence pour chaque variable explicative . . . 96
5.6 Odds Ratio des 5 variables les plus importantes de la régression logistique 97
5.7 Caractéristiques du modèle de forêt aléatoire sur les données d'apprentissage et de test . . . 98
5.8 Caractéristique du modèle de forêt aléatoire sur les données d'apprentissage et de test . . . 99
5.9 Scores d'importance . . . 101
5.10 Caractéristiques du réseau de neurones sur les données d'apprentissage et de test . . . 102
5.11 Caractéristiques du réseau de neurones sur les données d'apprentissage après suréchantillonage et les données de test . . . 103
5.12 Tableau du top 5 des variables les plus importantes selon le réseau de neurones . . . 104
5.13 Récapitulatif des résultats de nos modèles sur la base de test . . . 105

B.1 Suite des résultats du test de Khi2 (1/3) . . . 118
B.2 Suite des résultats du test de Khi2 (2/3) . . . 119
B.3 Suite des résultats du test de Khi2 (3/3) . . . 120

C.1 Tableau complet des odds Ratio . . . 122

D.1 Importance des variables pour le réseau de neurones . . . 124

Table des figures

1.1 Les différentes catégories des produits d'assurance (Source :[13]) 6
1.2 L'absentéisme aggravé par la crise de la Covid en France (Source : [18]) . 11

2.1 Illustration du fonctionnement de la méthode Bagging (Source : [7]) . . . 26
2.2 Précision sur le fonctionnement des réseaux de neurones (Source : [14]) . 29

3.1 Schématisation des jointures réalisées pour créer la BV CRM 35
3.2 Exemple de commentaire sur CRM avant traitement 39

4.1 Nombre de sinistres par rapport à la durée de l'arrêt 66
4.2 Nombre de sinistres par rapport à la cause du sinistre 67
4.3 Nombre de sinistres par rapport à l'âge de la survenance 67
4.4 Nombre de sinistres par rapport à la durée entre la survenance et la déclaration 68
4.5 Nombre de contrats pour chaque modalité de la sélection médicale 69
4.6 Nombre de contrats pour chaque modalité de la surprime 70
4.7 Nombre de contrats pour chaque modalité du top fumeur 71
4.8 Nombre de contrats avec un seul sinistre et avec plus d'un sinistre 72
4.9 Nombre de sinistres par cause de sinistre et sexe 73
4.10 Nombre de sinistres par âge de survenance et sexe 74
4.11 Nombre de sinistres par âge de survenance et cause 75
4.12 Nombre de sinistres par cause de sinistre et top fumeur 76
4.13 Nombre de sinistres par durée et décision 77
4.14 Nombre de sinistres par sélection médicale et décision 78
4.15 Nombre de sinistres par cause et décision 79
4.16 Nombre de sinistres par réseau de distribution et décision 79

5.1 Courbes ROC pour les données d'apprentissage et de test 91
5.2 Courbes ROC pour les données d'apprentissage et de test 93
5.3 Courbes ROC pour les données d'apprentissage suréchantillonnées et les données de test 95

5.4 Courbes ROC pour les données d'apprentissage et de test 98
5.5 Courbes ROC pour les données d'apprentissage après suréchantillonnage et les données de test 100
5.6 Courbes ROC pour les données d'apprentissage et de test 103
5.7 Courbes ROC pour les données d'apprentissage après suréchantillonnage et les données de test 104
5.8 Courbes ROC de nos 3 modèles sur les données de test 106

A.1 Première étape des jointures des données CRM aux données VANDA . . 114
A.2 Deuxième étape des jointures des données CRM aux données VANDA . . 115
A.3 Troisième étape des jointures des données CRM aux données VANDA . . 116

Bibliographie

[1] https ://www.universalis.fr/encyclopedie/alan-mathison-turing/4-intelligence-artificielle.

[2] https://datascientest.com/comment-la-data-science-est-utilisee-dans-le-monde-de-las

[3] https://www.thelem-assurances.fr/actualite-assurance/habitation/assurance-zone-inondable.

[4] Scénario : sélection de modèle par sélection de variables et pénalisation. *Institut de mathématiques de Toulouse*.

[5] Comment s'écrit l'intelligence artificielle ? lisp... sous la plume de john mc carthy. *Histoire CIGREF*, 2014.

[6] Data science : définition et enjeux. *Le pont*, 2022.

[7] Le bagging en machine learning, de quoi s'agit-il ? *Kobia Devançons l'avenir*, 2022.

[8] Scandale cambridge analytica : Facebook soumet un accord de principe. *La tribune*, 2022.

[9] Le règlement général sur la protection des données (rgpd), mode d'emploi. *Bercy Infos Entreprises*, 2023.

[10] Pierre Ailliot. Modèle linéaire, m1 euria. *M1 EURIA*, 2023.

[11] Yoboua ANGOUA. *Provisionnement Non-Vie*. Master 2 EURIA, 2022-2023.

[12] Charles Tremblay et Clément Côme. L'auc roc. *Kobia, Devançons l'avenir*, 2021.

[13] Alain Bouijoux et Gaëtan Vergely. Typologie des risques et leur impact pour un client particulier. *openclassrooms*, 2023.

[14] Christian Gagné. Perceptron multicouche. *Université LAVAL*.

[15] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow : concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[16] INSEE. Âge moyen de la mère à l'accouchement. 2023.

[17] WILL KENTON. What is the pearson coefficient ? definition, benefits, and history. *Investopedia*, 2022.

[18] Marion Kindermans. Les entreprises face au défi d'un absentéisme grandissant. *Les Echos*, 2021.

[19] Majid Joseph LAGNAOUI. Machine learning et nouvelle classe de problèmes actuariels : une illustration par l'étude de cas. *Institut des actuaires*, 2015.

[20] Paul MILAN. Algorithme sur la méthode newton-raphson. 2015.

[21] Anthony NAHELOU. Le risque arrêt de travail. *Cours EURIA - Master 1 – 2020 -2021*, 2021.

[22] Jean-Michel Poggi Robin Genuer. Arbres cart et forêts aléatoires, importance et sélection de variables. *HAL*, 2017.

[23] Noah Sdiri. L'absentéisme des salariés, un phénomène en forte hausse. *Capital*, 2022.

[24] Talend, a qlik company. *Quelles différences entre Data Science et Machine Learning ?*, 2023.

[25] José Unpingco. *Python for probability, statistics, and Machine Learning*. Springer Nature Switzerland AG, 2016.

[26] Cédric Villani. Blaise pascal par cédric villani : le jeu de la science et de la raison. *Philosophie magazine*, 2023.