Université Paris Dauphine
Master 2 Mathématiques, Apprentissage et Sciences humaines

HIÉRARCHISATION DES DÉTERMINANTS DES ARRÊTS MALADIE : ETUDE DE CAS DES ENTREPRISES ASSURÉES CHEZ MALAKOFF HUMANIS

Rédigé par :

Justin Kossivi AYIVI

Sous la supervision de :

Damien LOUREIRO
Études techniques - Malakoff Humanis

Sous l'encadrement de :
Mme Nadege FABRE
Malakoff Humanis
Études Techniques

Année académique 2021/2022

Table des matières

Remerciements

Je souhaite remercier toutes celles et ceux qui ont contribué de près ou de loin aux travaux présentés dans ce mémoire. Je remercie M. Guillaume SIMON, d'avoir accepté ma candidature pour cette offre de stage et de m'avoir permis de réaliser ce stage au sein de la direction support marchés et plus spécifiquement au sein de l'équipe études techniques.

Je souhaite remercier Mme Nadège FABRE, manager de l'équipe Études techniques, qui a bien voulu encadrer l'ensemble des travaux menant à la rédaction de ce mémoire. J'ai beaucoup apprécié travailler avec vous pendant ces cinq mois au courant duquel j'ai beaucoup appris sur l'assurance santé et la prévoyance. Merci pour le temps et les conseils que vous m'avez prodigués tout au long de cette rédaction.

Je souhaite remercier mes collègues de Malakoff Humanis qui n'ont ménagé aucun effort pour m'intégrer dans cette jeune équipe dynamique et épanouissante afin de permettre un déroulement dans de meilleures conditions du stage.

Je souhaite ensuite remercier particulièrement M. Damien LOUREIRO qui a fortement contribué aux travaux de ce mémoire et m'a permis par sa compréhension fine de la DSN de pouvoir explorer cette masse d'informations pour mon étude. Merci également à M. Bienvenu KENFACK NANDA à qui j'accorde une mention particulière pour m'avoir consacré beaucoup de son temps et de sa disponibilité sans faille.

Je remercie mon tuteur académique, M. Christophe DUTANG, pour sa relecture et ses conseils avisés.

Enfin, je tiens à remercier l'ensemble des enseignants du Master de Mathématiques, Apprentissage, Sciences et Humanités (MASH) de l'Université Paris Dauphine pour la qualité de l'enseignement.

Résumé

Une majorité des entreprises est aujourd'hui confrontée au problème d'absentéisme de leurs salariés. Cette situation amène les entreprises à investir des sommes non négligeables dans la compréhension des mécanismes qui sont à l'origine de ces arrêts. Si le terme absentéisme fait souvent référence à l'arrêt maladie (un salarié est absent parce qu'il déclare être en arrêt maladie) c'est surtout les déterminants de ces arrêts qui ont retenu notre attention dans le cadre de cette étude. Ce mémoire a pour objectif d'apporter une réponse à la question des déterminants de l'arrêt maladie et de la hiérarchisation de ces déterminants grâce aux algorithmes de machine learning. Cette approche orientée machine learning est justifiée par le succès obtenu par ces différents algorithmes face à des problèmes plus ou moins complexes allant de l'ordre linéaire jusqu'au non linéaire. Néanmoins, la garantie intrinsèque de ces algorithmes dépend fortement de la quantité et de la qualité des données disponibles. La déclaration Sociale Nominative (DSN) nous a offert la possibilité de disposer de cette masse d'informations, fiables et de très bonne qualité, pour étudier ces arrêts de travail dans le but de comprendre ces déterminants afin de mieux les anticiper. Dans un premier temps, ces travaux proposent une identification des déterminants des arrêts maladie à travers la littérature puis une analyse des mécanismes expliquant l'arrêt de travail chez le salarié. Dans un second temps, les algorithmes de machine learning tels que les arbres de décision CART, les méthodes d'agrégation de modèles tels que le Gradient Boosting, le XGBoost et le Light Gradient Boosting ont été utilisés pour modéliser notre variable d'intérêt qui est le nombre d'arrêts annuel par entreprise. Cette variable est un indicateur composite qui caractérise aussi bien la durée de l'arrêt, le nombre de salariés en arrêt ainsi que l'occurrence de l'arrêt. Cette modélisation a permis de conserver deux modèles importants à savoir le Gradient Boosting et le Light GBM qui obtiennent de bien meilleurs scores à l'issue de la phase de test. Enfin, l'étude a permis de ressortir, suivant le degré d'importance, les variables déterminantes dans la survenance du nombre annuel de jours d'absence dans une entreprise donnée mais de souligner également la nature de leur influence.

Mots clés : arrêt maladie, data science, Machine Learning, Arbres de décision CART, Bagging, Boosting, XGBoost, LightGBM, SHAP.

Première partie

Généralités et Contexte de l'étude

Chapitre 1

Généralités sur les arrêts de travail

Avant d'analyser les relations qui peuvent subsister entre l'arrêt de travail et les déterminants socio-économiques, il convient de comprendre le contexte dans lequel s'inscrit la survenance d'un arrêt de travail et le mode d'emploi utilisé par les organismes d'assurances pour proposer des garanties aux sinistrés en cas de survenance d'un arrêt. En France notamment, la législation permet de distinguer deux catégories d'assurance en cas de survenance d'arrêt : l'assurance incapacité et l'assurance invalidité. Ces deux notions font référence dans la majorité des cas au caractère temporel de l'arrêt notamment par sa durée, son caractère définitif ou non et bien d'autres critères. Mais avant de rentrer dans les détails, nous proposons dans un premier temps de donner les informations nécessaires à la compréhension des termes génériques lorsqu'on parle d'assurance incapacité et invalidité et dans un second temps d'aborder plus spécifiquement l'arrêt de travail.

1.1 Notions d'assurance incapacité et invalidité

Dans cette première section, nous définissons les termes récurrents dans le cas d'une assurance incapacité - invalidité.

Délai de carence : C'est la durée qui s'est écoulée entre la souscription d'un contrat d'assurance et le moment où la garantie liée à ce contrat prend effet. Un sinistre survenu pendant ce délai ne sera donc pas couvert par l'assurance (même s'il se prolonge au-delà du délai de carence). Il constitue un des éléments du contrat et la durée de sa période varie en fonction des organismes d'assurances et des garanties souhaitées par les assurés. Son but est d'éviter l'anti-sélection et de permettre dans une certaine mesure de dissuader les personnes qui veulent acheter une garantie en cas d'incapacité-invalidité parce qu'elles sont malades.

Franchise : La franchise est la période au delà de laquelle l'assuré ayant souscrit à un contrat d'assurance, perçoit des prestations suite à la survenance d'un sinistre. Sa durée est également un des éléments constitutifs du contrat. Par exemple, si un sinistre dure 30 jours et que la franchise retenue est de 10 jours, seulement 20 jours seront couverts par l'assurance. Une franchise de courte durée ainsi qu'un contrat sans franchise entraînerait des tarifs plus élevés et multiplierait des frais de gestion. Il permet, en ce sens, de pouvoir réduire l'anti-sélection.

Délai de rechute : Le délai de rechute est également un des éléments précisés dans le contrat. Si la durée entre la date de fin d'un premier sinistre et la date de survenance d'un second est inférieure ou égale au délai de rechute, et que le second sinistre a une cause liée au précédent, la franchise n'est pas réappliquée pour le second sinistre.

Concepts d'incapacité et d'invalidité Comme énoncé au début du chapitre, les notions d'assurance incapacité et invalidité relèvent du caractère temporaire ou non de la durée de l'arrêt.

Le concept de l'incapacité est synonyme d'*arrêt maladie*, ou encore *d'arrêt de travail* et sous-entend une interruption momentanée de travail. C'est dans cette catégorie que se retrouve un assuré quand il interrompt son travail pour cause de maladie ou d'accident. À ce moment, on ne connaît pas le moment où il reprendra le travail.

L'invalidité par contre requiert un caractère permanent c'est -à-dire un état d'invalidité irréversible. Néanmoins une distinction peut être faite entre les invalides qui ne peuvent plus occuper leur ancienne profession, de ceux qui ne peuvent plus en occuper aucune. Cette catégorisation permettra par la suite à l'assurance maladie de mieux spécifier les indemnisations que peuvent percevoir les assurés.

1.2 Les arrêts maladie

La raison subséquente d'un arrêt maladie est déjà évoquée dans l'énoncé : les salariés ont des arrêts parce qu'ils sont malades. Malheureusement, cette simple phrase ne suffit pas pour conclure une étude sur l'arrêt maladie car la

réalité s'avère bien plus compliquée. L'arrêt de travail n'est pas le résultat d'un processus totalement aléatoire, et de nombreux déterminants complexes peuvent l'expliquer et contribuer à l'anticiper.

1.2.1 Définition

Avant d'entrer dans le cœur du sujet, nous définissons ce que nous entendons par arrêts maladie. L'arrêt maladie est défini légalement comme une absence du travail pour raison d'accident ou de maladie non-professionnels qui doit être constatée par un médecin. Le salarié absent a 48 heures pour justifier son arrêt auprès de l'assurance maladie en transmettant un avis d'arrêt maladie auprès de son employeur et de l'assurance maladie. Nous distinguons d'autres types d'arrêt comme les congés parentaux ou les accidents du travail et maladies professionnelles. Ce mémoire n'abordera que le premier type d'arrêt car les données utilisées dans le cadre de l'étude ne reflètent que ce type d'arrêt.

En cas d'arrêt maladie prononcé par le médecin traitant, le salarié se voit suspendre son contrat de travail et percevra des indemnisations de la part des organismes d'assurance. La procédure d'indemnisation varie selon les salariés mais est, en général, partagée entre trois acteurs :

1. L'assurance maladie verse tout d'abord aux salariés des Indemnités Journalières (IJ) à hauteur de 50% du salaire journalier de base et dans la limite de 1,8 fois le SMIC après un délai de carence de trois jours pour la plupart des salariés du secteur privé (ce délai est d'un jour pour les salariés de la fonction publique). Ce régime est financé par les cotisations sociales obligatoires prélevées sur la paye de chaque salarié.
2. Un complément de cette indemnité est exigé de l'employeur en fonction des accords collectifs de branche et de sa convention collective. Au minimum, l'employeur doit assurer un certain niveau de salaire en cas d'arrêt de travail pour maladie ou accident aux salaires ayant au moins un an d'ancienneté au delà du huitième jour d'arrêt. Il doit compléter l'indemnité de base pour atteindre 90% du salaire de base. Ce complément est maintenu pendant le premier mois de l'arrêt puis est diminué progressivement.
3. Enfin, le salarié peut recourir aux assureurs complémentaires pour compléter l'indemnisation des arrêts après une certaine durée de franchise et combler le manque à gagner en cas d'incapacité / invalidité. Le montant et la durée de cette indemnisation vont dépendre des contrats signés entre le salarié et les organismes de protection sociale. Chez Malakoff Humanis, c'est l'entreprise qui souscrit cette couverture pour l'ensemble de ces salariés.

1.2.2 Conséquences des arrêts maladie

Nous identifions des conséquences positives et négatives des arrêts maladie. D'un point de vu positif, il permet d'accorder un temps de repos aux salariés afin de leur assurer un certain revenu lorsqu'ils ne sont pas en capacité de travailler. Cependant, cet arrêt peut engendrer des conséquences négatives que nous regroupons en trois principaux points :

— le salarié n'est pas toujours indemnisé à la hauteur de son salaire surtout lorsque l'arrêt est de durée longue. Ce qui peut conduire à un état de troubles psychologiques et/ou d'isolement,

— le salarié en arrêt maladie interrompt son activité et de ce fait fera peser sur son équipe les tâches qui lui ont été assignées dans ces fonctions. Cela pourrait entraîner un état de frustration pour le reste de l'équipe (car contrait d'ajouter à leurs tâches quotidiennes celles de leur collègue) et une baisse de la productivité de l'équipe,

— les arrêts maladie ont aussi des conséquences financières claires sur les acteurs chargés d'indemniser le salarié. Ces acteurs devront également faire face à des abus du système de protection sociale.

Tenter de prévenir les arrêts maladie aura ainsi un impact positif sur les finances des salariés, de l'État et des entreprises mais aussi sur la santé des salariés. Néanmoins, cette prévention ne devrait pas avoir l'objectif de contraindre les salariés de ne pas s'arrêter mais de prévenir les causes de ces arrêts puis qu'empêcher les salariés de prendre des arrêts peut mener à de possibles conséquences futures plus graves sur leur état de santé mentale comme psychologiques.

1.2.3 Causes des arrêts maladie

Les causes des arrêts maladie sont cependant diverses, variées et peuvent provenir de plusieurs sources. Les causes sont synthétisées dans le graphique 1. Ce graphe synthétique décrit les différents processus déterminant l'occurrence d'un arrêt maladie. Nous pouvons regrouper en deux catégories les déterminants de ces arrêts :

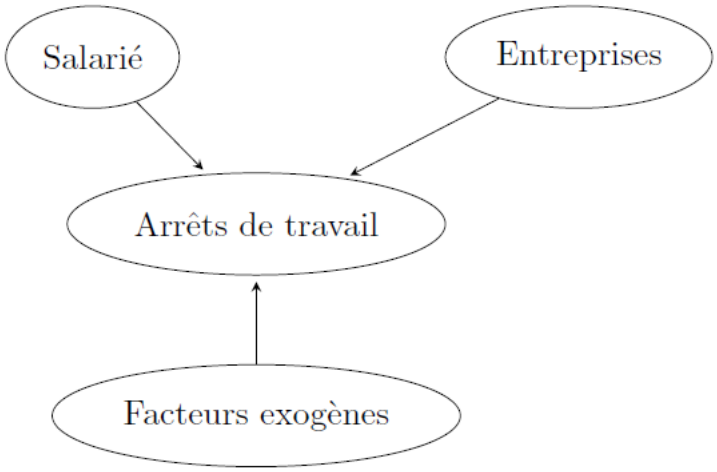


FIGURE 1 – Sources potentielles des arrêts maladie [19]

1. les déterminants endogènes relatifs aux salariés et aux entreprises tels que la santé mentale comme physique du salarié[1], l'hygiène de vie[2], son âge [3, 4], la pénibilité physique de l'emploi [5, 6], la pression exercée sur les salariés [7], le secteur d'activité, etc.
2. les déterminants exogènes, relatifs à tout le reste : maladies saisonnières [8], pandémie [9], les crises socio-économiques, etc.

1.2.4 Tarification

Au vu des différentes causes de l'arrêt maladie présentées à la section précédente, les assureurs doivent appliquer une segmentation tarifaire pour affiner au mieux le tarif des garanties. Elle dépend le plus souvent de :

- L'âge : c'est le premier paramètre qui entre en compte dans le calcul du tarif. Une personne âgée présente plus de risque d'être en arrêt maladie mais aussi d'avoir une durée de l'arrêt plus longue en moyenne,
- La catégorie socioprofessionnelle (CSP) : un assuré présentera plus ou moins de risque si son activité est manuelle ou sédentaire. Un assuré cadre aura en moyenne moins d'arrêts qu'un assuré non cadre,
- La zone géographique : Les assurés résidant en Corse, dans les DOM-TOM ou en région PACA auront des tarifs plus élevés car la propension à avoir un arrêt dans ces zones est d'autant plus grande.

D'autres paramètres peuvent influer sur le tarif comme la franchise désirée par l'assuré ou une éventuelle réduction de franchise en cas d'hospitalisation ou d'accident.

1.3 Terrain d'étude et enjeux actuels

1.3.1 Terrain d'étude

L'entreprise Malakoff Humanis est un groupe de protection sociale (c'est-à-dire un organisme qui met en œuvre des régimes de retraites complémentaires et des couvertures de protection sociale complémentaire), paritaire (piloté par les partenaires sociaux), mutualiste et à but non-lucratif. Malakoff Humanis est né suite au rapprochement entre Malakoff Médéric et Humanis, deux anciens groupes de protection sociale, le 1er janvier 2019.

Malakoff Humanis est concerné par la problématique des arrêts de travail puisqu'elle propose des contrats de prévoyance qui couvrent le risque d'incapacité pour les entreprises et les particuliers. L'incapacité d'un salarié est l'impossibilité temporaire à travailler (contrairement à l'invalidité, qui est une impossibilité permanente). Un salarié est dit en incapacité de travail lorsqu'il est en arrêt maladie pour une durée supérieure à une franchise déterminée par un contrat qui relie l'organisme de prévoyance et l'entreprise. L'organisme de prévoyance indemnise ainsi le salarié en

complément de l'assurance maladie et de l'entreprise. Outre la problématique d'assurance, Malakoff Humanis propose de nombreuses études et de nombreux services pour comprendre et maîtriser l'absentéisme des entreprises [19]. Le terme *absentéisme* fait référence ici à « toute absence qui aurait pu être évitée par une prévention suffisamment précoce des facteurs de dégradations des conditions de travail entendus au sens large : les ambiances physiques mais aussi l'organisation du travail, la qualité de la relation d'emploi, la conciliation des temps professionnel et privé, etc. » [19].

1.3.2 Enjeux actuels

Les enjeux de ce mémoire sont de deux ordres. Un premier enjeu serait d'identifier la connaissance des arrêts maladie dans les entreprises, le profilage des salariés absents et comment réduire leur risque d'absence en attachant une attention particulière aux déterminants des arrêts maladie. Un second enjeu serait de hiérarchiser ces différents déterminants pour identifier le profil d'entreprise susceptibles d'être le plus touché par les arrêts maladie et ainsi réadapter les garanties du contrat et donc rendre compte du niveau d'absentéisme dans les entreprises. Ce mémoire propose ainsi un premier volet d'analyse où le focus est mis sur les salariés et un deuxième niveau d'analyse se concentrera en détail sur les informations agrégées au niveau entreprise.

Les travaux présentés dans ce mémoire se basent principalement sur une source de données : **les Déclarations Sociales Nominatives (DSN)**. Elles décrivent l'ensemble des trajectoires d'absence des salariés répertoriant les informations les concernant et permettent ainsi de mettre en place des outils de suivi et de compréhension de l'absence au travail chez les entreprises.

Cette masse de données disponibles rend possible un ensemble d'études sur les arrêts de travail pour les assurés de Malakoff Humanis et devraient donc nourrir des réflexions générales autour de l'analyse des données d'arrêt maladie. L'axe d'étude que nous prenons renforce cette idée de proposer un outil de suivi de l'absentéisme des assurés de Malakoff Humanis. L'objectif des travaux est de caractériser l'absentéisme des entreprises et de sélectionner des informations personnalisées et pertinentes à partir des données mises à disposition.

1.4 Objectif des travaux et axes d'analyse

1.4.1 Objectif des travaux

L'objectif principal de ce mémoire est de développer une approche comparative de surveillance des arrêts maladie des entreprises. Dans ce mémoire, la surveillance est entendu au sens large et désigne la description des facteurs qui définissent un phénomène d'intérêt (les arrêts maladie dans notre cas). Pour répondre à cet objectif, nous construisons pour chaque entreprise, une valeur repère afin d'évaluer son niveau d'absentéisme pour chacun des indicateurs socio-démographiques. Un des objectifs spécifiques est d'identifier l'ensemble de ces indicateurs et ensuite de procéder à une mise en application des modèles de **data science** sur ces données.

1.4.2 Plan du mémoire

Dans un premier temps, l'objectif sera d'analyser les facteurs qui entrent en jeu dans la survenance d'un arrêt maladie : quels sont les déterminants de ces arrêts ? Quels influences pourraient-ils avoir sur les arrêts, Ces arrêts sont -ils plus spécifiques à certaines entreprises ? Pour aborder le problème, nous commencerons par une revue de littérature des différents déterminants de l'arrêt maladie puis illustrerons également les méthodes statistiques couramment utilisées pour les analyser. Ensuite, il sera question de mettre en exergue le processus qui a conduit à la construction de la base de données agrégée au niveau entreprise.

Nous terminerons par une analyse des limites et perspectives.
Le langage de programmation utilisé dans le chapitre 2 est R.
Le langage de programmation utilisé à partir du chapitre 3 est Python.

Deuxième partie

Cadre méthodologique

Chapitre 2

Construction de la base

Comme dans tout projet de data-science, les données sont le cœur du sujet. L'objectif de ce chapitre est d'arriver à fournir une base de données fiable et en quantité suffisante afin de s'assurer la robustesse et la précision des résultats de la modélisation. Comme énoncé dans le premier chapitre, la base de données utilisée dans le cadre de cette étude est issue des bases de la déclaration sociale nominative. Elle a contribué à extraire une série d'informations en procédant à une jointure de plusieurs bases disponibles dans la DSN. Avant de rentrer plus en détail sur la base, nous proposons de présenter, dans un premier temps, la Déclaration Sociale Nominative puis dans un deuxième temps d'aborder plus spécifiquement les déterminants des arrêts maladie et enfin de présenter la typologie de variables retenues et les critères de construction de la base finale.

2.1 Sources des données : La Déclaration Sociale Nominative

La Déclaration Sociale Nominative (DSN) est une déclaration en ligne, exigée de tout employeur du secteur privé depuis 2017. Elle permet de déclarer de façon unique, en ligne et tous les mois, un ensemble d'informations concernant les salariés. Elle est directement produite à partir de la fiche de paie du salarié et renseigne les données concernant la paie du salarié ainsi que les évènements concernant les périodes d'activité du salarié tels que le salaire, les cotisations payées aux différents organismes, le numéro SIRET (pour « système d'identification du répertoire des établissements ») de l'établissement, les numéros de contrats de prévoyance et santé complémentaire, l'arrêt de travail, la paternité etc. La DSN est effectuée par établissement (SIRET), c'est-à-dire qu'une entreprise enverra autant de DSN qu'elle compte d'établissements. La DSN a deux principales fonctions : l'une étant de calculer et de payer les cotisations sociales et l'autre d'informer automatiquement tous les organismes sociaux tels que Pôle emploi, Assurance maladie (CPAM : Caisse primaire d'assurance maladie), Urssaf, Agirc-Arrco (Association générale des institutions de retraite des cadres - Association pour le régime de retraite complémentaire des salariés), organismes complémentaires de santé (mutuelle, assurance ou institution de prévoyance), des données concernant les salariés (leurs rémunérations, leurs activités, etc.). Par ailleurs, la DSN sert également de canal utilisé par l'Etat dans le cadre du prélèvement à la source mis en place depuis le 1er janvier 2019. Elle sert notamment à indiquer les taux à appliquer aux salariés et à transmettre les paiements à la Direction Générale des Finances Publiques.

La DSN est très riche en information puisqu'elle remplace une multitude d'autres déclarations (Urssaf, Pôle emploi, CPAM, AGIRC-ARRCO, Organismes de prévoyance, etc.). Toutes ces données sont réparties en blocs qui correspondent chacun à un thème précis. "Déclaration", "Entreprise", "Individu" ou "Fin de contrat" en sont des exemples. En tant qu'organisme de prévoyance, Malakoff Humanis reçoit les blocs liés à son activité d'assurance, et ce, uniquement pour ses entreprises clientes. Les institutions de Prévoyance ne visualisent que les blocs les concernant. Par exemple, Malakoff Humanis visualisera le bloc "Affiliation Prévoyance" et "Arrêt de travail" alors que les blocs "Actions gratuites" ou "Options sur titres (stock options)" ne lui seront pas disponibles. Néanmoins, juste avec les bases à disposition, nous disposons d'une quantité suffisante d'informations et uniquement certaines données ont été requises pour notre étude. Afin de construire notre base de données, une sélection des différents blocs s'est révélée nécessaire afin d'unifier les informations importantes pour l'étude, dispersées dans différents blocs. Cette préparation de données a été faite par d'autres personnes de l'équipe et antérieurement aux travaux du-dit mémoire. Cependant, même si les blocs ont été retraités pour une thématique particulière comme le datamart [1] "Arrêt de travail", un deuxième niveau de retraitement a été effectué. Ensuite, les différents **datamarts** utilisés sont les datamarts "déclaration", "arrêt de travail", "effectif entreprise". L'ensemble de ces datamarts sont antérieurs à cette étude et sont notamment utilisés pour d'autres sujets autre que celui de ce mémoire. Même si la DSN présente de nombreux avantages, quelques inconvénients sont aussi à mentionner :

Avantages et inconvénients de la DSN

1. Un datamart est un ensemble de données relatives au même thème. L'idée d'un datamart est de présenter les données de manière organisée et souvent agrégée pour répondre à un besoin spécifique, ce qui permet de faciliter l'usage de la donnée en entreprise.

Voici quelques avantages que les acteurs pourraient tirer de l'utilisation de la Déclaration Sociale Nominative :

1. La DSN assure un certain niveau de qualité et de fiabilité des données du fait qu'elle relève d'une démarche administrative exigée de tout employeur.
2. Le fait que la déclaration soit mensuelle réduit le risque d'erreur/omission et permet de sécuriser et de fiabiliser des obligations sociales avec moins de contentieux et de pénalités.
3. Elle réduit le nombre de déclarations à effectuer, ce qui permet de simplifier une importante partie des démarches administratives des employeurs auprès des différents organismes (CPAM, Urssaf, etc.).
4. Elle permet également la disponibilité quasi-immédiate des informations actualisées sur les salariées et sur les sinistres. Ainsi, l'information sur la prise d'un arrêt de travail par un salarié est directement identifiable le mois suivant dans la DSN alors que ces arrêts ne seront vus dans nos bases de gestion qu'après le délai de franchise et de gestion.
5. Enfin, la DSN facilite l'accès à une immense quantité de données. De nombreuses informations sont centralisées via le même canal (salaires, arrêts de travail, affiliations, contrat etc) alors que ces diverses informations dépendaient jadis de canaux différents rendant difficile leur accès.

Comme le souligne [20], l'utilisation de la DSN présente aussi quelques inconvénients. Cette masse d'informations rendu disponible et accessible dans la DSN est un réel challenge pour les assureurs de personnes car elle permettra de procéder à différentes études. Néanmoins, les données ne sont pas facilement exploitables puisqu'elles sont souvent réparties en différents blocs et pas forcément à une maille intéressante pour les actuaires et les *data scientists*. C'est pourquoi l'utilisation des datamarts "DM_DSN" facilite considérablement l'utilisation de ces données. Ensuite, même si les assureurs utilisent un logiciel de paie compatible en DSN, des erreurs de déclaration peuvent être présentes du fait que la DSN [2] est relativement récente et que beaucoup d'entreprises ne maîtrisent pas encore totalement l'outil ou qu'il n'y avait pas de référent dédié à la gestion de la paie notamment pour les entreprises de moins de 50 salariés.

Néanmoins, après la déclaration des entreprises, il est possible de procéder à des corrections en cas d'anomalie ou de non conformité des déclarations émises par le biais d'un outil développé par l'Urssaf appelé 'Suivi DSN'. Il est ainsi garanti une certaine qualité et fiabilité dans les DSN.

2.2 Justification des variables explicatives

Nous rappelons que l'objectif principal est de répondre à la question de l'hiérarchisation des déterminants de l'arrêt maladie. Une première question importante est d'identifier les déterminants plausibles de l'arrêt maladie. Dans la littérature, nous en avons une multitude ([1, 2, 3, 4, 5, 6, 7, 8, 9]), mais nous conservons les variables les plus citées que nous détaillerons par la suite :

Âge : Elle est sans doute l'une des variables les plus explicatives de la survenance d'un arrêt chez un individu. Comme abordé succinctement dans la section 1.2.4, elle constitue l'un des critères de segmentation de la tarification des contrats d'assurance prévoyance et santé. Il est vraisemblable qu'un individu plus âgé ait plus de chance d'avoir un arrêt qu'un jeune.

CSP : La catégorie socioprofessionnelle peut influencer également sur la sinistralité. En effet, on peut supposer qu'un cadre aura moins tendance à se mettre en arrêt maladie qu'un ouvrier par exemple, ce dernier exerçant des tâches plus « pénibles ».

Sexe : Le sexe est également un critère important. Selon les résultats obtenus par le baromètre annuel bien-être et Santé au travail [3], les femmes ont un plus grand risque d'avoir un arrêt long et fréquent.

2. Elle a été mise en place en 2017

3. Le baromètre Bien-être et Santé au travail (BST) est une enquête annuelle dont l'objectif est d'évaluer année après année le bien-être et la santé au travail des salariés. Elle est reconduite chaque année et interroge tous les ans environ 3500 salariés représentatifs du secteur privé français et c'est une enquête mise en place par Malakoff Humanis.

Type d'emploi : Selon que le type d'emploi occupé par le salarié soit à temps plein ou à temps partiel aura en moyenne une incidence sur la prise d'arrêt. En effet, pour un salarié occupant un emploi à temps partiel, on pourrait imaginer une baisse de la rémunération par rapport à un salarié ayant des qualifications équivalentes mais à temps plein sur le même poste. Cette situation pourrait emmener le salarié à temps partiel à se questionner davantage sur la prise d'un arrêt car le manque à gagner sera énorme.

Nature du contrat : Un contrat à durée indéterminé (CDI) évoque chez le salarié une situation financière stable et à long terme. Ainsi plus ou moins d'arrêts seront acceptés après une certaine période d'ancienneté, chose que ne peut se permettre facilement un salarié ayant un contrat à durée déterminée (CDD) ou sous contrat par intérim.

Taille de l'entreprise : Elle est une fonction croissante du nombre annuel de jours d'absence. La taille d'entreprise influe positivement sur le nombre d'arrêts. Cette variable nous paraît importante dans la mesure où elle permet de rapporter la fréquence et la durée de l'arrêt à l'effectif des salariés d'une entreprise donnée pour obtenir une estimation de son niveau d'absentéisme global.

Secteur d'activité : Il est également un des facteurs à prendre en compte dans la caractérisation des déterminants des arrêts maladie. Selon que le secteur d'activité soit le commerce, le BTP, l'industrie, la Santé, le social, les services, les transports, télécom et industrie, l'occurrence des arrêts et surtout leur durée n'est pas du tout la même.

On doute bien que ces déterminants ne suffisent pas pour expliquer un niveau d'absentéisme dans les entreprises et que d'autres déterminants, [19],

— d'ordre personnels comme :
  - la santé mentale et physique
  - l'hygiène de vie
  - l'équilibre de la vie personnelle

— d'ordre professionnels comme :
  - la pénibilité physique
  - l'activité professionnelle

— d'ordre purement exogènes
  - les épidemies, les pandémies
  - les phénomènes météréologiques
  - les chocs socio-économiques (faillite, chute boursière, crise économique, etc)
  - les grèves

pourraient également contribuer à expliquer la prise d'un arrêt et de sa durée. Néanmoins, vu l'impossibilité d'extraire la majorité de ces informations des bases de données DSN mises à notre disposition, nous avions orienté l'analyse vers les premiers déterminants indiquant les caractéristiques socio-démographiques des arrêts de travail.

2.3 Traitements des données

Comme évoqué précédemment, un datamart unifié contenant les principales informations de la DSN, nommé "DM_DSN" a été créé lors de précédents travaux. À partir de ce dernier, un deuxième datamart centré sur les arrêts de travail et un troisième datamart centré sur les effectifs des entreprises par mois et par année ont été construits pour des projets antérieurs à l'étude. Le premier datamart nommé "DM_DSN" englobe l'ensemble des déclarations mensuelles d'un établissement pour l'ensemble de ces salariés. Il retrace la trajectoire d'absence de chacun des salariés mais aussi contient des informations diverses faisant référence à cet arrêt comme la date de début d'arrêt, la date de reprise du travail, les périodes de ces différentes déclarations, etc. Cependant, lorsqu'un salarié n'a pas pris d'arrêt, l'information est rendue aussi disponible dans le "DM_DSN". C'est justement pour cela que le deuxième datamart "ArrT DSN" est intéressant car il permet d'obtenir une ligne par arrêt de travail de chaque salarié en agrégeant les différentes informations par rapport à la durée de l'arrêt, l'occurrence de l'arrêt (un second arrêt survenu à l'intérieur du délai de rechute est immédiatement rattaché au premier arrêt) et la période de survenance de l'arrêt. Les variables de déclaration telles que le premier et le dernier moment de connaissance de l'arrêt sont ainsi tous renseignés dans le

datamart "ArrT DSN" alors que les informations reflétant les caractéristiques socio-démographiques des salariés sont disponibles dans le datamart initial "DM_DSN".

Pour constituer la base finale, nous avions sélectionné, par échantillonnage aléatoire, un périmètre de 3039 entreprises. La segmentation par marchés des entreprises en trois types sur la base du troisième datamart "Effectif DSN" que sont : TPE-PE (effectif de l'entreprise de moins de 10 salariés), ME-ETI (effectif compris entre 10 et 1000 salariés) et GC (plus de 1000 salariés), nous a permis de procéder à un échantillonnage stratifié en conservant, dans la base finale, une proportion de 4% pour les GC, 72% pour les ME-ETI et de 24% pour les TPE-PE. Ainsi à partir de cette sélection aléatoire, nous avions considéré l'ensemble des individus relevant de ces différentes entreprises. Nous soulignons que c'est la maille entreprise , i.e SIREN, qui a été considérée et non SIRET i.e Etablissement pour simplifier le processus d'échantillonage. La partie suivante a été de créer une base d'arrêt de travail dans un premier temps en considérant les individus ayant au moins un arrêt dans la base de datamart DSN (car dans cette base, certaines déclarations mensuelles ne faisaient pas l'objet d'arrêt) puis de ne conserver que les entreprises ayant conclu des contrat à temps plein et à temps partiel avec leur salariés et dont la nature du contrat est indéterminée sur l'année 2021 uniquement.

Le choix de cette année pour l'étude peut être problématique mais elle nous paraissait être la moins polluée de phénomènes socio-économiques telles que la pandémie de la covid apparue en fin 2019 et dont les effets pervers s'étendent jusqu'en début 2022. Les anciennes déclarations de La DSN (année 2017 et 2018) faisaient notamment l'objet de beaucoup d'incohérences et nécessiterait un effort conséquent pour les retraitements. L'année 2022 étant en cours, il serait difficile de se prononcer sur la continuité de la dynamique des arrêts jusqu'en fin d'année.

En ce qui concerne les types de contrats, on pourrait également privilégier d'autres types de contrats de nature variée mais ces variables ne nous intéressent pas dans le cadre de cette étude. Ensuite, trois jointures ont été faites entre les datamart "DM_DSN" et "ArrT DSN" et "Effectif DSN" pour ne conserver que les informations sur les variables telles que l'âge, le sexe, la nature du contrat, le type du contrat, la date du dernier jour de travail (la date de prise de l'arrêt), la date de reprise de travail, la taille d'entreprise, le nombre jours d'arrêts, l'occurence d'arrêt, le mois de déclaration de l'arrêt, etc.

Cela nous a conduit à obtenir un échantillon de 1 087 208 individus et de 26 variables. Nous rappelons que dans cette table chaque ligne représente un arrêt de travail pour un salarié d'une entreprise donnée. Un salarié peut bien avoir deux ou plusieurs lignes dans cette base représentant chacune une périodicité différente de l'arrêt. L'objectif de la section suivante sera de faire une brève description du jeu de données sur laquelle la modélisation a été effectuée.

2.4 Typologie d'entreprises

Cette section est essentielle afin de mieux appréhender la structure des données avant le processus de modélisation. Il est question ici de présenter la typologie des variables qui serviront à modéliser la variable cible et quelques analyses préliminaires pour réduire au mieux la dimensionnalité (nombre) des variables explicatives. L'objectif de ce travail est de construire, pour chaque entreprise, une valeur repère afin d'évaluer son niveau d'absentéisme pour chacun de ses indicateurs. L'indicateur sélectionné est le nombre annuel de jours d'absence selon les variables sociodémographiques. Nous avons sélectionné le nombre de jour d'absence pour représenter l'absentéisme globale puisqu'il s'agit d'un indicateur composite, décrivant aussi bien la proportion de salariés absents que la durée des arrêts, ce qui représente assez bien le niveau d'absentéisme global d'une entreprise. La base non agrégée ainsi obtenue présente les statistiques descriptives consignées dans la table 1. Elles ont servi à constituer la base d'agrégation qui ne contenait qu'au final 3039 individus i.e SIREN et dont les différents indicateurs ont été agrégés en proportions.

2.5 Description des variables

En considérant la base de départ énoncé au chapitre 2, nous avons agrégé les données au niveau annuel et au niveau de l'entreprise. La variable à expliquer est **le nombre annuel de jours d'absence**. Elle représente la somme du nombre de jours arrêtés pour tous les salariés dans une entreprise donnée au cours de l'année 2021. Les variables suivantes sont utilisées pour l'expliquer :

- Proportion de salariés de moins de 35 ans, entre 35 et 44 ans, de 45 à 54 ans, de plus de 55 ans en 2021,

Variable explicative	Effectif	Proportions (%)
Tranche d'âge	674101	100.0
Entre 35 et 44 ans	214769	31.9
Entre 45 et 54 ans	181482	26.9
Moins de 35 ans	163362	24.2
Plus de 55 ans	114488	17.0
Sexe	674101	100.0
Féminin	285212	42.3
Masculin	295451	43.8
Non mentionne	93438	13.9
Type de cdi	674101	100.0
cdi autre	249	0.0
cdi prive	671319	99.6
cdi public	2533	0.4
Type emploi	674101	100.0
Temps partiel	87009	12.9
Temps plein	587092	87.1
csp	674101	100.0
Cadre	159185	23.6
Employé	416298	61.8
Indépendant	35	0.0
Ouvrier	98583	14.6

TABLE 1 – Tableau synthétique sur les variables explicatives

Pour des raisons de confidentialité, un coefficient multiplicatif est appliqué sur les chiffres présentés dans le tableau 1.

- Proportion de cadres, de professions intermédiaires, d'ouvriers et d'employés en 2021
- Proportion de salariés à temps plein et à temps partiel en 2021,
- Proportion de salariés de sexe masculin et féminin en 2021,
- Proportion de CDI dans le privé, dans le public et dans d'autres secteurs que le public/privé,
- Effectif des salariés par SIREN.

Les variables sont toutes ainsi de type quantitative et ne disposent d'aucune valeur manquante.

Chapitre 3

Modélisation

L'objet de ce chapitre est de présenter les concepts théoriques utiles à la compréhension des modèles prédictifs qui seront utilisés dans la suite du mémoire. Nous soulignons tout d'abord que l'objectif est d'évaluer le niveau d'absentéisme global d'une entreprise. Pour cela, trois critères sont nécessaires pour juger de ce niveau d'absentéisme : la proportion de salariés en arrêt maladie, la durée moyenne de ces arrêts et le nombre de jours d'arrêt maladie. L'objectif premier de ce mémoire est d'appliquer un modèle de Data Science plus spécifiquement d'apprentissage automatique (*machine learning*) afin d'évaluer leur comportement face à ce type de données. Nous rappelons que le principe de base d'un modèle d'apprentissage supervisé est de déceler le lien entre des données d'entrée appelées *input* et les données de sortie appelées *output*. Pour parvenir à cette fin, le modèle de machine learning utilise une fonction de coût qu'il faudra minimiser par le biais d'un algorithme d'optimisation. Un modèle ayant de bonnes performances est celui en effet capable de trouver les bonnes sorties aux données d'entrée qui lui seront fournis. Nous avions à cet effet considéré trois types de modèles à explorer essentiellement ceux basés sur les arbres de décisions.

3.1 Les arbres de décision

Très populaire dans le domaine de la data science, les arbres de décision sont un modèle d'apprentissage automatique adapté aussi bien aux tâches de régression et de classification. Elles sont capables de performer sur des tâches complexes notamment en rapport avec des données non linéaires. Nous apportons quelques détails à leur fonctionnement, comment ils sont entraînés et par quels processus se font leurs prédictions. Tout d'abord les arbres de décision se définissent comme étant une structure en forme d'organigramme qui permettent de classer des données d'entrée en fonction des données en sortie. La construction de l'arbre se fait en quatre grandes étapes :

- ✧ On commence par choisir la variable séparant le mieux les individus de chaque classe en fonction de la variable cible, en sous-populations appelées nœuds : le critère précis de choix de la variable et de sa valeur testée dépend de chaque type d'arbre,
- ✧ Pour chaque nœud, on répète la même opération, ce qui donne naissance à un ou plusieurs nœuds fils. Chaque nœud fils donne à son tour naissance à un ou plusieurs nœuds, et ainsi de suite, jusque ce que :
- ✧ La séparation des individus ne soit plus possible,
- ✧ L'un des critères d'arrêt d'approfondissement de l'arbre soit satisfait.

Nous avons trois principaux arbres de décision :

- CHAID (CHi-Square Automation Interaction Detection) qui utilise le test du χ^2 pour définir la variable la plus significative et le découpage de ses modalités. Il est donc plus adapté à l'étude des variables explicatives discrètes,
- CART (Classification and Regression Tree) utilise l'indice de Gini pour maximiser la pureté des nœuds (on y reviendra en détail dans la section suivante). Il est adapté à l'étude de tout type de variables explicatives,
- C5.0 cherche à maximiser le gain d'information réalisé en affectant chaque individu à une branche de l'arbre. Il est adapté à l'étude de tout type de variables explicatives.

La différence principale de ces arbres réside dans la mesure de sélection d'un attribut ou plus précisément d'un critère de subdivision ou de découpage de l'arbre. Nous nous concentrerons sur le type CART qui est implémenté dans la bibliothèque *Scikit-learn* et qui nous a servis à tester toutes sortes de modèles.

3.1.1 Arbre de type CART

D'une façon précise, CART est en effet l'algorithme d'optimisation utilisée par Scikit-learn pour entraîner des arbres de décision. Comme énoncé dans la section précédente,Scikit-Learn utilise l'algorithme CART (Classification And Regression Tree) pour former des arbres de décision (également appelés arbres "croissants"). L'idée est vraiment très simple : l'algorithme divise d'abord l'ensemble d'apprentissage en deux sous-ensembles à l'aide d'une seule

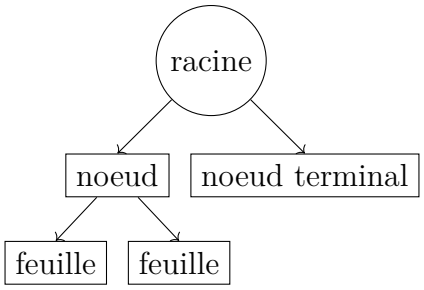


FIGURE 2 – Illustration simplifiée d'un arbre CART

caractéristique k et d'un seuil t_k. Comment choisit-il k et t_k ? Il recherche la paire (k, t_k) qui produit les sous-ensembles les plus purs (pondérés par leur taille). La mesure de **la pureté du noeud** se fait par l'évaluation de la fonction de coût associé à ce noeud. Plus elle possède une valeur faible, mieux la subdivision nous apporte le maximum d'informations. La fonction de coût qu'il convient de minimiser dans une tâche de régression est donnée par l'équation suivante :

$$J(k, t_k) = \frac{m_{\text{gauche}}}{m}\text{MSE}_{\text{gauche}} + \frac{m_{\text{droite}}}{m}\text{MSE}_{\text{droite}} \text{ avec } \begin{cases} \text{MSE}_{\text{noeud}} = \sum\limits_{i \in \text{noeud}} (\hat{y}_{\text{noeud}} - y^{(i)})^2 \\ \hat{y}_{\text{noeud}} = \frac{1}{m_{\text{noeud}}} \sum\limits_{i \in \text{noeud}} y^{(i)} \end{cases}$$

$m_{\text{gauche/droite}}$ étant le nombre d'instances dans le sous-ensemble gauche/droite de l'arbre et $\text{MSE}_{\text{gauche/droite}}$ désignant l'erreur quadratique moyenne dans le sous-ensemble gauche/droite de l'arbre.

NB : *L'algorithme de type CART (confer figure 2) disponible dans Scikit-learn ne produit que des arbres binaires c'est à dire qu'un noeud parent ne peut avoir plus de deux noeuds fils. Évidemment, il existe d'autres algorithmes qui produisent ce type d'arbre à l'image des ID3 [10].*

Une fois, cette division effectuée les sous ensembles seront subdivisés en sous ensembles de sous ensembles et ainsi de suite jusqu'à ce qu'un critère d'approfondissement soit atteint. Ce critère peut être la profondeur maximale de l'arbre défini avant la phase d'entraînement ou la subdivision pour laquelle l'impureté ne diminue plus significativement ou encore d'autres hyperparamètres résultant de la structure de l'arbre.

3.1.2 Régularisation des hyperparamètres

Malgré la simplicité des arbres de décision du point de vue de l'interprétabilité, ils présentent un inconvénient majeur qui est le **surapprentissage**. Il désigne la capacité d'un modèle à prédire parfaitement les données sur lesquelles il a été entraîné mais dont aucune généralisation n'est possible. C'est-à dire qu'aucune extrapolation de nos données ne sera possible avec ce type de modèle. Pour pallier ce problème du au surapprentissage, nous disposons de quelques boites d'outils dont la régularisation. Elle permet en effet de lutter contre le sur paramétrage du modèle en le restreignant à un nombre de paramètres suffisant. Il peut s'agir, par exemple, de réduire la profondeur maximale (*max_depth*) de l'arbre, qui est définie sur une valeur infinie par défaut ou de réduire le nombre maximal de caractéristiques (*max_features*) avant la subdivision au niveau de chaque noeud. L'autre astuce qui prévient contre le surapprentissage est **l'élagage**. Cette opération est effectuée après la constitution de l'arbre sans aucune restriction et consiste à supprimer les sous arbres qui n'améliorent pas significativement l'erreur globale du modèle ou plus spécifiquement l'impureté du noeud. Le processus est réitéré jusqu'à ce que aucun noeud non informatif (qui apporte une information pertinente) ne soit identifiable. A l'opposé, la régularisation pourrait également passer par l'augmentation des valeurs des hyperparamètres tels que le nombre minimal d'échantillons par feuille (valant 1 par défaut, *min_samples_leaf = 1*) ou la taille minimale de l'échantillon requis pour diviser un noeud (qui vaut 2 par défaut *min_samples_split = 2*). Afin de trouver le nombre optimal d'arbres, nous pouvons aussi utiliser l'arrêt anticipé (*early stopping*, en anglais).

3.1.3 Instabilité des arbres de décision

Les arbres de décision sont simples à comprendre et à interpréter, faciles à utiliser, polyvalents et puissants. Cependant, ils sont très sensibles aux petites variations qui peuvent figurer dans les données affectant tout autant leur généralisation même si nous appliquons une régularisation, comme présenté à la section précédente. En effet, il suffit de modifier une ou deux structures de variables dans la structure des données de base pour qu'on ait deux arbres complètement différents. Une manière de réduire cette instabilité due aux arbres de décision est de limiter le caractère aléatoire des données présentées à l'algorithme. Les forêts aléatoires constituent l'un des algorithmes qui permettent de réduire l'effet de cet aléa. Ils combinent plusieurs arbres aléatoires en moyennant l'ensemble de leurs prédictions. Ces types de modèles sont en général dans la catégorie des modèles ensemblistes (plus connu en anglais sous *Ensemble Learning*). On en distingue trois types : le Bagging, le Boosting et le Stacking.

3.2 Présentation des modèles d'Ensemble Learning

Conceptuellement, la technique d'Ensemble Learning repose sur un concept assez simple. Imaginons que l'on veuille trouver une réponse à une question et qu'on a le choix de la poser à un expert directement ou à une multitude d'individus pris au hasard. Il y a beaucoup de chances que les réponses multiples des individus agrégées permettent de mieux répondre à la question qu'au seul avis de l'expert. C'est sur ce principe que se base les méthodes d'ensemble. Ainsi, si l'on regroupe les prédictions d'un groupe de prédicteurs (tels que des classificateurs ou des régresseurs), nous obtiendrons globalement de meilleures prédictions qu'avec le meilleur prédicteur individuel. Mais certaines conditions doivent être réunies :

- ✧ une taille suffisamment grande des données d'entraînement,
- ✧ chaque modèle apporte un minimum d'informations,
- ✧ les predicteurs soient indépendants les uns des autres,
- ✧ les modèles sont diversifiés.

3.2.1 Bagging

Cette méthode tient son nom de la concaténation de *Boostrap et Agging*. Elle repose sur une technique d'échantillonage avec remise. C'est à dire qu'on procède à un tirage aléatoire avec remise pour constituer différentes portions aléatoires du jeu de données sur lesquelles le même algorithme, est appliqué. L'intérêt de cette technique est qu'elle permet de réduire la variabilité dans les prédictions due au fait que l'arbre individuel se spécialise assez rapidement sur les données d'entraînement. Lorsu'on combine un ensemble d'arbres entraînés sur des portions aléatoires bootstrapées du jeu de données pour obtenir des prédicteurs performants et moins sujets au surapprentissage, l'on parle des **forêts aléatoires** (*Random Forest* en anglais). L'algorithme Random Forest introduit un caractère aléatoire supplémentaire lors de la croissance des arbres. Au lieu de rechercher la meilleure caractéristique lors de la division d'un nœud, il recherche la meilleure caractéristique parmi un sous-ensemble aléatoire de caractéristiques. Il en résulte une plus grande diversité d'arbres qui satisfait un certain compromis entre le biais et la variance. Ce qui permets d'obtenir en général un meilleur modèle.

3.2.2 Boosting

Comme l'algorithme Random Forest, la technique du Boosting est une technique d'apprentissage automatique basée sur l'assemblage de modèles de prédictions faibles, généralement les arbres de décision. Elle permet d'améliorer tout modèle d'apprentissage automatique en l'entraînant de manière itérative à la résolution des points faibles des modèles précedents. Parmi la multitude de méthode de boosting disponibles, les plus populaires sont *AdaBoost* et *Gradient Boosting*. La technique du Gradient Boosting a été préférée à l'AdaBoost dans le cadre de ce mémoire. En effet, le Gradient Boosting permet d'ajuster à chaque itération, lors de la phase d'entraînement, les prédicteurs aux résidus au lieu de pondérer conformément aux erreurs produites par le modèle comme c'est le cas avec l'AdaBoost. Il en résulte un modèle qui s'adapte mieux aux résidus du modèle contrairement à l'Adaboost qui se concentre sur les erreurs individuelles. Une variante du Gradient Boosting est le *XGBoost* pour *Extreme Gradient Boosting*. Elle est une version optimisée du Gradient Boosting en ce sens qu'il est paréllisable et donc permet une efficacité d'exécution de l'algorithme. Il offre également la possibilité d'entraîner ce type de modèle sur un algorithme de base comme les modèles linéaires contrairement au Gradient Boosting qui ne s'opère que sur les arbres de décision. En parlant

d'efficacité d'exécution nous disposons également d'une version du Gradient Boosting dévéloppé par Microsoft du nom de Light Gradient Boosting Machine ou plus simplement **Light GBM** qui a montré sur un ensemble de jeu de données de très grande taille une exécution beaucoup plus rapide et une meilleure précision que le XGBoost.

3.2.3 Le stacking

Le bagging et le boosting combinaient l'ensemble des prédicteurs pour construire un prédicteur final afin améliorer la précision globale. Pour des tâches de classification, le critère d'agrégat qu'est utilisé est le mode c'est à dire que la prédiciton finale est la prédiction la plus fréquente parmi l'ensemble des prédicteurs alors que pour les tâches de régression le critère d'aggrégat est plutôt la moyenne. L'idée du stacking est d'entraîner le modèle afin de trouver le meilleur ajustement possible ou la meilleure fonction qui fait correspondre les valeurs prédites par les prédicteurs aux valeurs de sortie.

3.3 Évaluation de la performance d'un modèle

Avant de rentrer un peu plus en détail sur cette section, il convient de distinguer deux principales approches pour évaluer de la performance d'un modèle, un premier point est de mesurer le comportement intrinsèque de l'algorithme notamment sur les données d'apprentissage (training set) et les données de test (testing set) ; la seconde approche consiste à retenir un algorithme au vu de ces indicateurs de qualité puis à le comparer à un ensemble d'autres algorithmes en utilisant des métriques de comparaison. Ces deux étapes s'entremêlent. Pour ce qui est de la première étape, nous disposons de plusieurs critères mais la qualité d'un modèle dépend bien souvent de deux notions importantes en apprentissage automatique que sont : le sous-apprentissage et le surapprentissage. Elles ont en commun la notion de compromis biais-variance bien connu en machine learning. La figure 3 résume bien à quoi s'apparente ce phénomène.

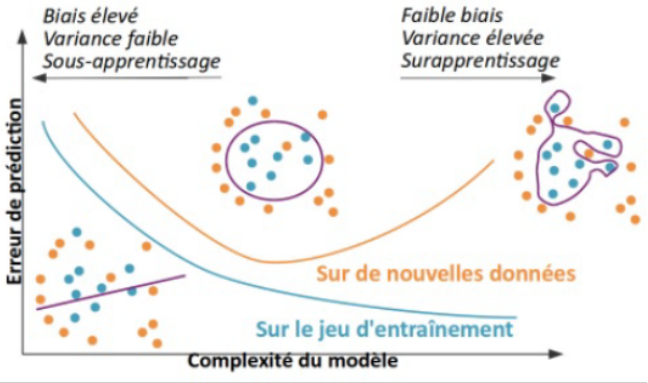


FIGURE 3 – Lien entre complexité et erreur de prédiction

Ainsi, un modèle beaucoup trop complexe c'est-a-dire avec un nombre significativement élevé de paramètres (ou de variables) s'apparente plus à du surapprentissage alors que le sous-apprentissage résulte plus d'un modèle trop simple c-a-d avec peu de variables et qui souffre intensément d'un problème de biais. De façon sommaire, les caractéristiques d'un modèle en surapprentissage sont les suivantes :

- ♠ biais faible
- ♠ variance elevée

tandis qu'un modèle en sous-apprentissage affiche l'inverse de ces caractéristiques. Le meilleur modèle serait celui qui affiche une variance faible et un biais faible. Ce qui reste impossible puisque ces deux expressions varient en sens opposé. En pratique, le mieux est de réduire le biais en gardant un niveau de variance acceptable.

3.3.1 Validation croisée

Une des techniques qui permet de lutter contre le sur-apprentissage est la validation croisée. Un type de validation croisée est le *k-fold cross validation* en anglais. Elle consiste à découper une base de données en k segments disjoints de taille égale : k-1 servent à apprendre le modèle et 1 segment est utilisé lors de la phase de test. L'opération est répétée aléatoirement k fois pour que chaque segment soit utilisé une fois comme test. Finalement, l'erreur est calculée comme la moyenne des erreurs de chacun des k modèles. Schématiquement, nous avions la figure 4 pour une base d'apprentissage de 70% et une base de test de 30%. Par ailleurs, cette phase est regulièrement accompagnée par le réglage d'hyperparamètres du modèle (*fine tuning*). Comme énoncé dans la section 2 sur la régularisation, les hyperparamètres peuvent influer sur la capacité du modèle à surapprendre. Ainsi, en couplant la recherche des meilleurs valeurs pour les hyperparamètres ainsi que de la validation croisée, le modèle retenu sera beaucoup plus robuste et aura moins tendance à surapprendre. Afin de les déterminer, deux méthodes de recherche peuvent être utilisées :

- ☆ La méthode Grid Search : C'est la méthode d'optimisation classique qui permet de tester une série de paramètres dont les valeurs sont définies à l'avance. L'algorithme teste alors chaque combinaison possible. La limite de cette méthode est l'explosion des temps de calcul si le nombre de combinaisons de valeurs à tester est trop grand. Exemple : dans le cas d'un Random Forest, on cherche le paramètre optimal relatif au nombre d'arbres $ntrees \in \{100, 45, 30\}$
- ☆ La méthode Random Search : A la différence de la méthode précédente, les paramètres optimaux ne sont pas recherchés exhaustivement au sein d'un nombre fini de valeurs testées mais aléatoirement dans un intervalle de valeurs. Suivant la largeur de la plage de valeurs, les temps de calcul peuvent être longs. Par exemple, dans le cas d'un Random Forest, on cherche le paramètre optimal relatif au nombre d'arbres $ntrees \in [100, 500]$.

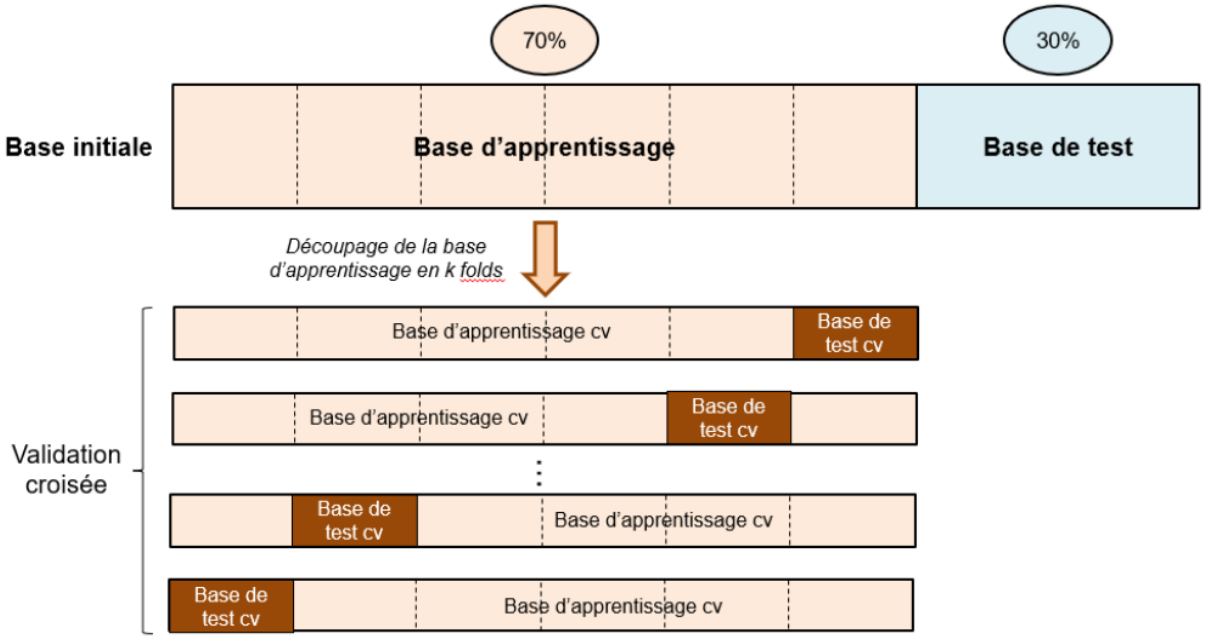


FIGURE 4 – Principe de validation croisée

Source : [21], page 66

3.3.2 Métriques et évaluations des régressions

Dans les sections précédentes, on évoquait le terme d'erreur qui est un indicateur de la qualité globale. Si l'erreur quadratique moyenne (*mean squared error*, en anglais) demeure la plus connue pour les tâches de régression, d'autres indicateurs de qualité de régression comme le MAE ou encore le R^2 sont souvent utilisées. Cette section a pour but de présenter les indicateurs ou plus précisément les métriques qui permettent de juger de la qualité d'un modèle par

rapport à un autre. Dans sa forme mathématique, il est la somme des carrés des résidus normalisée par le nombre d'observations ie :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

où n est le nombre d'observations, y_i est la valeur réelle de la cible que nous essayons de prédire pour l'observation i, et $\hat{y}_i$ est la valeur prédite par le modèle pour y_i. Pour se ramener à l'unité de y, on peut prendre la racine de la MSE : on obtient ainsi la RMSE, ou Root Mean Squared Error :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Ces indicateurs présentent l'avantage de pénaliser plus fortement les fortes erreurs (à travers le carré) que d'autres mesures de performance.

Mais la RMSE ne se comporte pas très bien quand les étiquettes peuvent prendre des valeurs qui s'étalent sur plusieurs ordres de grandeurs. Imaginons faire une erreur de 100 unités sur une étiquette qui vaut 4, le terme correspondant dans la RMSE vaut $100^2 = 10\ 000$. C'est exactement la même chose que de faire une erreur de 100 unités sur une étiquette qui vaut 8000. Cependant une prédiction de 104 au lieu de 4 (soit une erreur de 2 ordres de grandeur) paraît être une erreur bien plus grande qu'une prédiction de 8100 au lieu de 8000.

Pour prendre cela en compte, on peut passer les valeurs prédites et les vraies valeurs au log avant de calculer la RMSE. On obtient ainsi la RMSLE (Root Mean Squared Log Error) dont l'expression est :

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(y_i + 1) - log(\hat{y}_i + 1))^2}$$

Une métrique alternative courante d'évaluation de la régression est le R^2 qui mesure la quantité de la variabilité expliquée par le modèle :

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Néanmoins il a le défaut d'augmenter mécaniquement si l'on ajoute des variables explicatives même non pertinentes dans notre modèle. Son interprétation ne doit pas être hâtive et surtout appuyée par d'autres métriques d'évaluation.

Nous avons également l'erreur moyenne absolue (*mean absolute erreur*, en anglais). Elle exprime la moyenne de l'erreur de prédiction du modèle absolue. Si le modèle était parfait, la valeur serait égale à zéro, mais il n'y a pas de limite supérieure à cette métrique, à la différence du coefficient de détermination. Elle est en revanche facile à interpréter parce qu'elle est exprimée en unités de la cible. C'est la métrique à utiliser si l'on a un faible intérêt pour la prédiction des valeurs aberrantes.

$$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i|$$

Nous avons également l'erreur absolue en pourcentage moyenne (MAPE), également connue sous le nom d'écart absolu moyen en pourcentage (MAPD) qui a la particularité d'être sensible aux erreurs relatives. Elle n'est, par exemple, pas modifiée par une mise à l'échelle globale de la variable cible. Elle s'écrit :

$$\text{MAPE}(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}\frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)}$$

où ϵ est un nombre arbitraire petit mais strictement positif pour éviter des résultats indéfinis lorsque y est zéro.

Une autre métrique également intéressante est l'erreur absolue médian. Elle est notamment robuste contre l'influence des outliers (des observations pour la variable y qui s'éloignent significativement de la tendance des autres observations). Elle s'écrit :

$$\text{MedAE}(y, \hat{y}) = median(|y_1 - \hat{y_1}|, \cdots, |y_n - \hat{y_n}|).$$

Une notion importante est de considérer que chacune de ces métriques considère différemment les erreurs produites par le modèle. Ainsi, suivant les données sur lesquelles le modèle sera appliqué et l'attention que l'on accordera au signal d'erreur des différents modèles, on pourra procéder à un choix. Est ce que l'on se concentre plus sur les erreurs importantes ou sur les valeurs aberrantes ? !

Diagramme des résidus : Une autre manière d'évaluer la qualité d'un modèle est de tracer son diagramme de résidus. Ce diagramme montre la différence entre les résidus sur l'axe vertical et la variable dépendante sur l'axe horizontal, ce qui permet de détecter les régions au sein de la variable cible qui peuvent être sujettes à plus ou moins d'erreurs. De ce fait, il est possible de montrer graphiquement la présence des aberrants et l'impact qu'ils peuvent avoir sur la qualité d'ajustement.

Une utilisation courante du graphique des résidus consiste à analyser la variance de l'erreur du régresseur. Si les points sont dispersés de manière aléatoire autour de l'axe horizontal, un modèle de régression linéaire est généralement approprié pour les données ; sinon, un modèle non linéaire est plus approprié. Nous pouvons également voir sur l'histogramme la distribution normale ou anormale de l'erreur autour de zéro.

3.4 Explication des modèles de régression

Cette section relève d'une notion fondamentale qu'est l'explicabilité des modèles de machine learning. Elle va de pair avec la notion d'interprétabilité. En effet, on dit d'un modèle qu'il est interprétable s'il est possible de formuler mathématiquement le fonctionnement de son algorithme mais qu'il est difficile de l'expliquer en des termes simples pour un utilisateur néophyte ou technophile. Alors qu'un modèle est explicable s'il est possible de formuler en des termes simples le procédé utilisé pour obtenir des prédictions. Les modèles interprétables s'apparentent plus à des modèles de type boites noires (*black boxes*) comme les forêts aléatoires et les réseaux de neurones etc alors que les modèles explicables relèvent de la catégorie des modèles de types boites blanches (ex : les arbres de décision, la régression linéaire).

C'est dans cet ordre d'idées qu'un intérêt particulier est souvent accordé à l'explicabilité des modèles complexes car il contribue à renforcer la confiance humaine dans la prédiction des modèles. Pour ce faire, on procède d'une manière simple : on se concentre davantage sur les interprétations locales qui expliquent les prédictions individuelles. En résumé, un modèle d'explication locale est un modèle linéaire (comme souvent plus facile à expliquer) que nous entraînons pour qu'il génère des prédictions aussi proches que possible du modèle original (modèle clé) autour d'une entrée.

À cette fin, plusieurs techniques ont été proposées dans la littérature existante. LIME et SHAP sont deux approches populaires d'explication locale, agnostiques vis-à-vis des modèles, conçues pour expliquer n'importe quel modèle de boîte noire donné. Ces méthodes expliquent les prédictions individuelles de tout classificateur d'une manière interprétable et fidèle, en apprenant un modèle interprétable (par exemple, un modèle linéaire) localement autour de chaque prédiction. Plus précisément, LIME et SHAP estiment les attributions des caractéristiques sur les instances individuelles, qui capturent la contribution de chaque caractéristique sur la prédiction de la boîte noire. Nous fournissons quelques détails sur ces approches, tout en soulignant comment elles sont liées les unes aux autres.

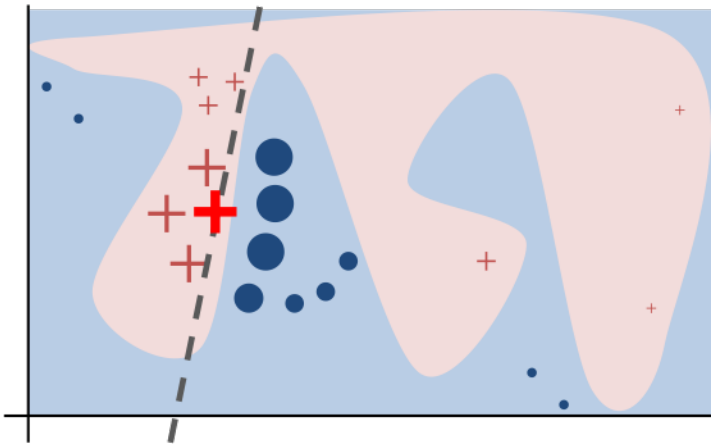
3.4.1 Introduction à LIME

LIME [4] a été publié pour la première fois en 2016 par Ribeiro, Singh et Guestri[18]. L'idée de LIME est d'expliquer une prédiction en remplaçant le modèle complexe par un modèle substitut interprétable localement.Il faut procéder en 3 étapes :

- ⋆ Échantillonner autour des données d'entrée (c-à-d du voisinage) et obtenir les prédictions correspondantes,

4. LIME signifie *Local Interpretable Model-agnostic Explanations*

FIGURE 5 – illustration de la méthode LIME



Source : https://github.com/marcotcr/lime

- ☆ Calculer les poids qui mesurent la proximité d'un point d'échantillonnage par rapport à l'entrée originale. Plus il est proche, plus il est grand,
- ☆ Entraîner une régression linéaire qui minimise la distance pondérée entre le modèle linéaire et les prédictions de l'échantillon.

De façon résumée, l'approche LIME permet pour un point de donnée ou un échantillon de découvrir les caractéristiques ont été prépondérantes dans le résultat de la prédiction. La technique consiste à perturber le point de données que nous voulons expliquer en créant un modèle linéaire local autour de ce point comme son explication (confer figure 5). Dans cette figure, La croix rouge vif représente la prédiction à expliquer et les points de couleur rouge non vifs proviennent de la perturbation du point de données vif, la ligne en pointillé représente ainsi l'approximation du modèle linéaire local aux points de données et servira d'explication du point de donnée initial.

Néanmoins, cette méthode comporte des inconvénients majeurs. La définition correcte du voisinage est un problème très important et non résolu lorsque l'on utilise LIME avec des données tabulaires. La complexité du modèle d'explication doit être définie à l'avance.

Il y a la possibilité de l'instabilité des explications. Dans la référence [23], les auteurs ont montré que les explications de deux points très proches variaient si l'on répète le processus d'échantillonnage, les explications qui en ressortent peuvent être différentes. L'instabilité signifie qu'il est difficile de faire confiance aux explications et qu'il faut être très critique. Les explications LIME peuvent être manipulées par le spécialiste des données pour cacher les biais [24]. La possibilité de manipulation rend plus difficile la confiance dans les explications générées avec LIME.

Les modèles de substitution locaux, avec LIME sont très prometteurs. Mais la méthode est encore en phase de développement et de nombreux problèmes doivent être résolus avant de pouvoir l'appliquer en toute sérénité.

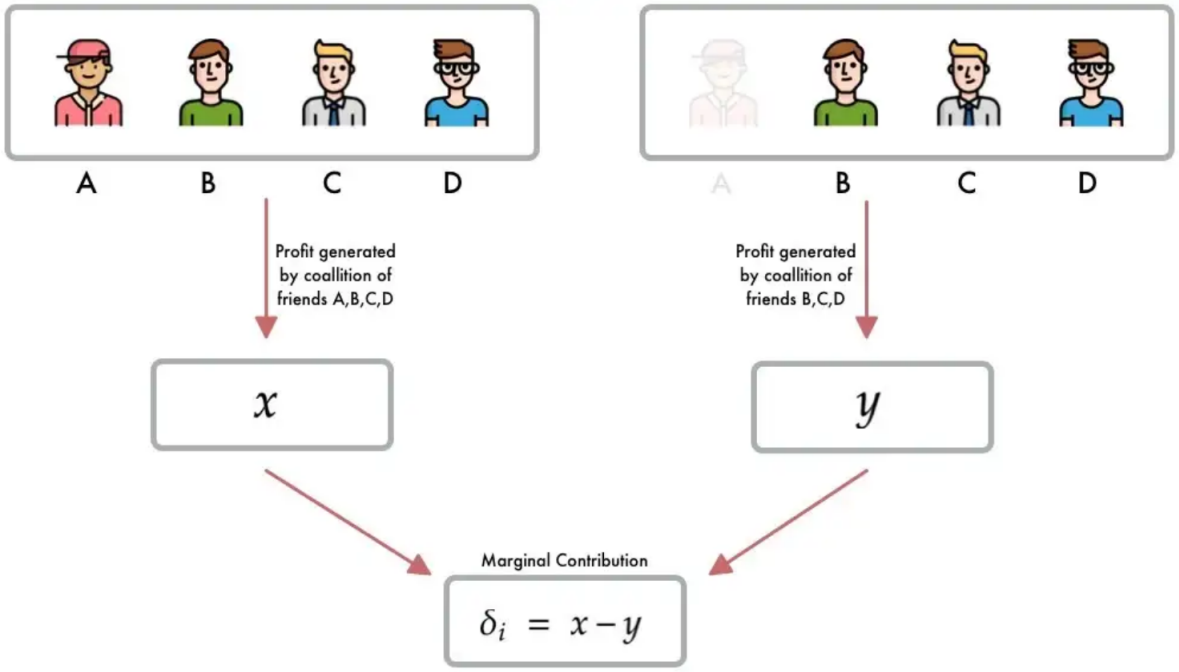
3.4.2 Introduction à SHAP

SHAP [5] a été publié pour la première fois en 2017 par Lundberg et Lee [17]. Pour comprendre l'idée de SHAP, nous devons d'abord savoir ce que sont les valeurs de Shapley.
La valeur de Shapley provient de la théorie des jeux. Il s'agit de la contribution marginale moyenne attendue d'un joueur dans toutes les combinaisons possibles au succès dans un jeu d'équipe.

Le graphique 6 fournit une très belle illustration graphique. Disons que quatre personnes forment une équipe pour réaliser un bénéfice. Le joueur A est identifié par le chapeau rouge qu'il porte. Pour comprendre quelle est sa contribution, nous devons calculer la différence de bénéfices lorsque le "chapeau rouge" fait partie de l'équipe et lorsqu'il n'en fait pas partie, selon 4 scénarios :

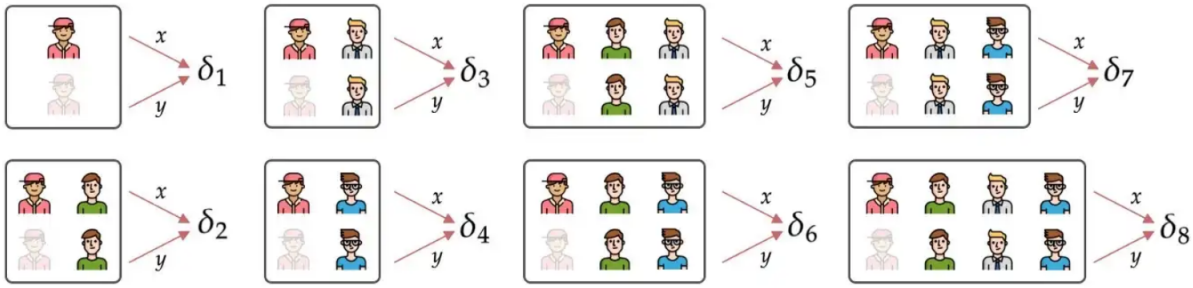
5. SHAP signifie *SHapley Additive exPlanations*

FIGURE 6 – Contribution marginale du joueur A à la coalition B, C, D



Source : [16]

FIGURE 7 – Calcul de la valeur Shapley pour le joueur A



The Shapley value for member

is given by:

$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$

Source : [16]

— l'équipe ne compte qu'une seule personne,

— l'équipe compte 2 personnes,

— l'équipe compte 3 personnes,

— l'équipe compte 4 personnes.

La contribution du "chapeau rouge" est la différence moyenne des profits pour toutes les combinaisons d'équipes (c'est-à-dire les "coalitions") dans ces 4 scénarios, c'est-à-dire la "valeur de Shapley". LA figure 7 en est une illustration. Cependant, lorsque le modèle original est compliqué (avec de nombreuses caractéristiques), le calcul de la valeur de Shapley est très intensif. C'est pourquoi la valeur SHAP est introduite.

La logique de la valeur de Shapley est simple, mais son effort de calcul peut devenir exponentiel lorsque le nombre de coalitions augmente. Pour pallier ce problème, au lieu de faire une boucle sur toutes les coalitions, SHAP utilise l'échantillon de coalition dans le même esprit que LIME :

— On commence par échantillonner au sein de différentes coalitions,

— Ensuite, on calcule les poids des différentes coalitions,

— Enfin, on entraîne le modèle linéaire qui minimise la distance pondérée entre le modèle linéaire et les prédictions de l'échantillon.

Les coefficients linéaires finaux constituent les valeurs SHAP.

3.5 D'autres méthodes d'explication des modèles

En dehors de LIME et SHAP, nous avions d'autres méthodes telles que la mesure d'importance des caractéristiques (*feature importance* en anglais) ou encore de la mesure d'importance de l'interaction entre les caractéristiques (*feature interaction*) qui peuvent servir également à des fins d'explicabilité des modèles. L'importance des caractéristiques, en particulier, sert à inspecter la façon dont les caractéristiques impactent un modèle mais ne nous nous permet pas de savoir dans quelle mesure cet impact varie en fonction du changement de la valeur de la variable. Elle démeure néanmoins toujours utilisée pour une hiérarchisation de l'impact global des différentes variables dans le modèle.

Troisième partie

Application à la hiérarchisation des déterminants de l'arrêt maladie

Chapitre 4

Présentation des résultats et discussions

Le but de ce chapitre est d'appliquer l'ensemble des algorithmes précédemment illustrés à la base de données. Nous avions vu au chapitre 1 le contexte méthodologique dans lequel s'inscrivait ce mémoire. Nous avions ensuite parcouru la typologie des différentes variables qui ont retenu notre attention et comment elles ont été traitées dans le cadre de notre étude. Puis, dans le chapitre 3, les différentes techniques et modèles de machine learning ont été présentés. Nous arrivons ainsi au but de cette étude qui est de la prédictibilité des arrêts en fonction des facteurs socio-économiques mais en particulier de la hiérarchisation de ces différentes variables dans la contribution à l'explicabilité du nombre annuel de jours d'absence.

4.1 Traitements préliminaires

Nous avons fait recours à un agrégat pouvant caractérisé globalement les différents facteurs socio-économiques dans une entreprise. Le choix a été ainsi tourné vers une agrégation des différentes variables telles que le sexe, l'âge, la catégorie socio-professionnelle, le type d'emploi (temps partiel, temps plein) ou le profil d'emploi (CDI privé/public). Tous ces déterminants ont été agrégés par proportions au niveau de la maille entreprise.

Après la constitution de cette nouvelle base "agrégée" qui résume par ligne les informations correspondantes à chaque entreprise suivant les différents déterminants, un premier niveau de retraitement a été fait. Nous avons supprimé de la nouvelle base les variables jugées non discriminantes. Il s'agissait du contrat de travail de type CDI dans le privé. En effet, la plupart des salariés dans l'échantillon obtenu étaient sous contrat CDI dans le privé.

Après cette phase de retraitement, nous avons sélectionné un ensemble de modèles de machine learning tels que les arbres de décision, les méthodes ensemblistes du type Boosting basés sur le Gradient Boosting Machine à savoir le XGBoost et le LightGBM. Afin d'évaluer les performances réelles de nos modèles, nous suivons deux étapes clés dans le processus d'apprentissage.

En premier lieu, nous subdivisons notre base de données en trois sous ensembles : une base d'apprentissage contenant 60% de l'ensemble des données (*training set*), une base de validation contenant 20% du reste des données (*validation set*) puis une base de test contenant les 20% restants (*testing set*). La base de test ne sera utilisée qu'en dernier ressort après un réglage judicieux des hyperparamètres et une prise en compte des signaux d'erreurs du modèle sur la base de validation. Nous suivons ces étapes pour l'ensemble des modèles illustrés ci-dessus :

- Entraînement du modèle sans hyperparamétrages
- Recherche des meilleurs hyperparamétrages du modèle par la méthode de GridSearchCV et du RandomSearchCV
- Entraînement du modèle avec ces hyperparamètres
- Phase de test du modèle par les Mesures et métriques de performances
- Tracé des différents diagrammes servant d'interprétation dont l'importance des caractéristiques

En deuxième lieu, sur la base des signaux retournés par les différents algorithmes, nous procédons à un choix de modèles puis à l'explicabilité de ces modèles par des méthodes du type SHAP values mais aussi de l'importance des caractéristiques. Enfin, la dernière partie reviendra à tester une fois le modèle retenu sur la base de test.

4.2 Transformation du nombre annuel de jours d'absence par siren

Le nombre annuel de jours d'absence est une variable quantitative et varie fortement en fonction de la taille d'entreprise. Conscient que sa variabilité pourra influer sur la capacité prédictive de nos différents modèles, nous optons pour une transformation de la variable à expliquer. On pourrait se demander les avantages à tirer de cette opération. Un premier avantage de cette pratique est l'accélération de l'apprentissage car en imposant à la variable cible une étendue de valeur plus faible, nous facilitons globalement la phase d'apprentissage du modèle rendant ainsi notre modèle plus précis et plus robuste face à l'influence potentielle des outliers. Cette transformation s'accompagne le plus souvent par une hausse des métriques tels que le R^2 ou une baisse des métriques tels que le MAE.

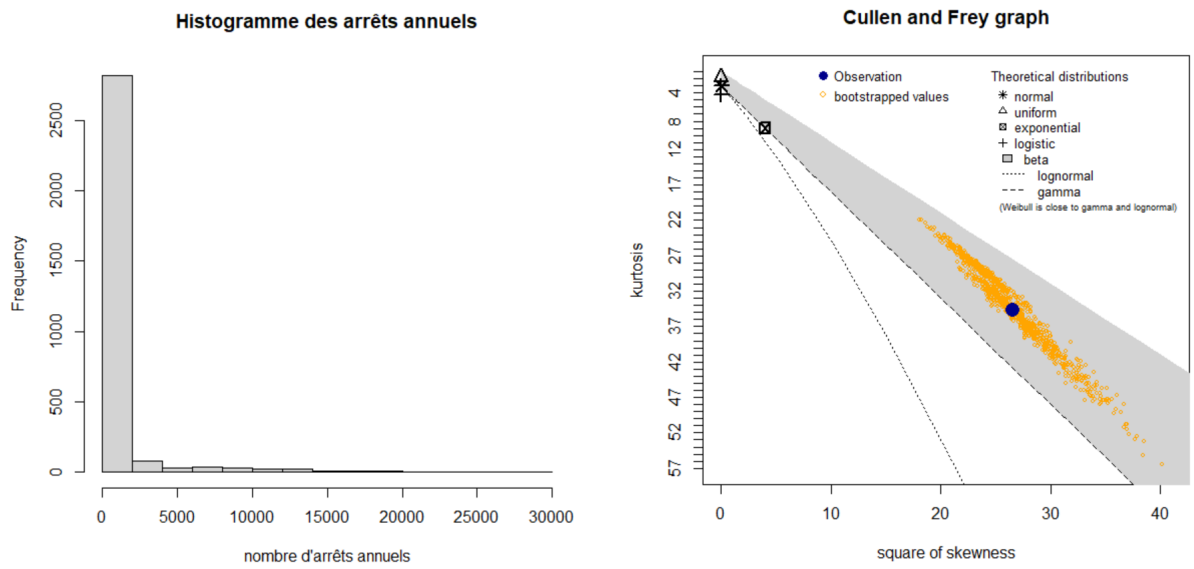


FIGURE 8 – Histogramme des arrêts annuels et graphe de Cullen et Frey

Il faudrait trouver une loi de distribution à laquelle pourrait s'apparenter la distribution de la variable cible. Pour mener cette réflexion, nous avions en premier lieu tracer l'histogramme du nombre annuel de jours d'absence (désignant y) et identifier, du moins graphiquement, la distribution de loi celle qui s'apparentait le plus à cette donnée. En utilisant la librairie **fidistrplus**[6] de R, nous obtenons le graphique 8 (à gauche, l'histogramme des arrêts annuels et à droite le graphe de Cullen et Frey permettant de savoir la loi qui s'ajuste au mieux à notre variable cible).

Nous avons identifié que la loi de distribution de type gamma semblait bien s'ajuster au nombre d'arrêts. Ce qui nous semble cohérent vu que notre variable exprime la durée d'un évènement, en l'occurrence, l'arrêt de travail. On peut s'arrêter à l'unique interprétation de ce graphique mais cela demeurerait insuffisant pour juger de la qualité de l'ajustement. la librairie **fidistrplus** nous propose une série de graphes pour confirmer de l'adéquation à cette loi. En considérant la loi gamma comme un bon fit, nous tombons sur le graphe 9. Ce graphique nous permet de conclure à un ajustement relativement moyen de la loi gamma en observant le graphe des P-P plot et des Q-Q plot. Une estimation de ces coefficients a été possible mais avec un intervalle de confiance très large rendant incertains la confiance dans les estimateurs.

La deuxième approche a été de tenter l'ajustement à une loi normale. Mais le graphe de qualité d'ajustement était moins intéressant. La dernière approche utilisée est celui de la transformation de la variable cible en utilisant l'information contenue dans les différents quantiles (les quantiles sont les valeurs qui divisent un jeu de données en intervalles de probabilité égale).

Cette méthode transforme les caractéristiques pour qu'elles suivent une distribution uniforme ou normale. Par conséquent, pour une caractéristique donnée, cette transformation tend à répartir les valeurs les plus fréquentes. Elle réduit également l'impact des valeurs aberrantes (marginales).

Pour pouvoir prédire la valeur correspondante de la cible, la transformation est appliquée à chaque variable indépendamment. Tout d'abord, une estimation de la fonction de distribution cumulative d'une variable est utilisée pour mettre en correspondance les valeurs originales avec une distribution uniforme. Les valeurs obtenues sont ensuite

6. Voici les principales options de cette boite à outils :
   - descdist : fournit un graphique de skewness-kurtosis pour aider à choisir le(s) meilleur(s) candidat(s) pour ajuster un ensemble de données
   - fitdist et plot.fitdist : pour une distribution donnée, estiment les paramètres et fournissent des graphiques de qualité d'ajustement.
   - bootdist : pour une distribution adaptée, simule l'incertitude des paramètres estimés par rééchantillonnage (bootstrap)

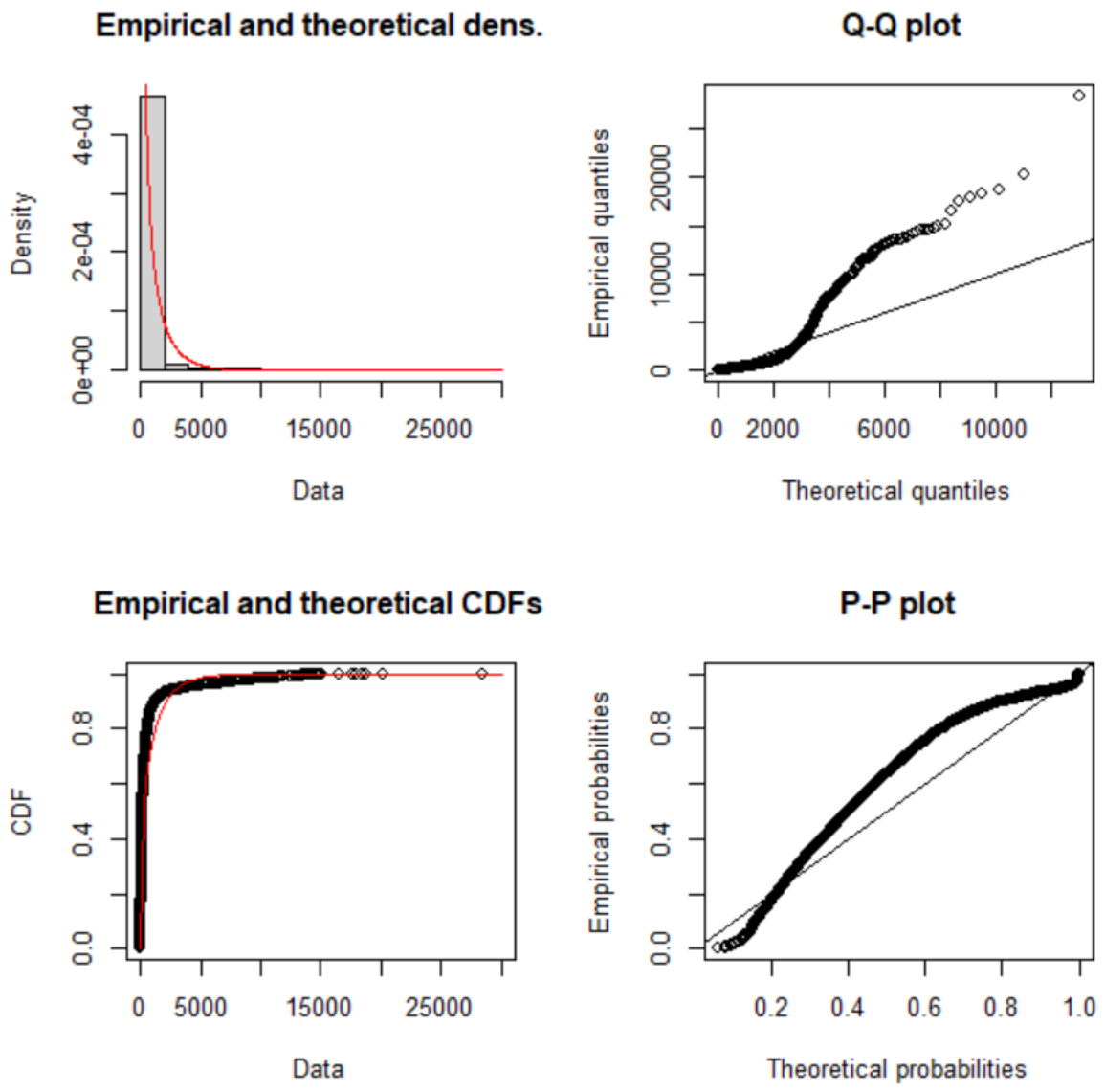


FIGURE 9 – Graphes de qualité d'ajustement

mises en correspondance avec la distribution de sortie souhaitée en utilisant la fonction quantile associée. Les valeurs des variables des données nouvelles/non vues qui se situent en dessous ou au-dessus de la plage ajustée seront mises en correspondance avec les limites de la distribution de sortie. Cette transformation est non linéaire. Elle peut fausser les corrélations linéaires entre les variables mesurées à la même échelle mais rend les variables mesurées à des échelles différentes plus directement comparables.

Cette technique nous a été utile pour appliquer une transformation non linéaire au nombre annuel de jours d'absence. Nous la désignons par y dans la suite du chapitre. Cette transformation est rendue possible grâce au module *quantileTransformer*[7] de *Scikit-Learn*. En guise de comparaison, pour ces différents modèles, nous avons testé le modèle sans et avec la transformation appliquée sur la variable d'intérêt.

4.3 Modélisation du nombre d'arrêts annuel

Nous nous intéressons dans ce qui suit aux résultats produits par les modèles sur la modélisation du nombre annuel de jours d'absence. Nous rappelons que quatre algorithmes ont retenu notre attention à savoir les arbres de décision, le gradient boosting (GBM) et deux (02) de ses variantes XGBoost et LightGBM. Les différents algorithmes ont été appliqués avec les valeurs par défaut afin d'observer les performances avant de procéder à une optimisation des hyperparamètres dans la quête de meilleures performances. Pour chaque modèle, une comparaison au préalable a été effectuée entre le modèle ayant subi une transformation et le modèle sans transformation de la variable d'intérêt.

4.3.1 Arbres de décision du type CART

L'arbre construit nous permet d'obtenir la valeur moyenne du nombre annuel de jours d'absence sur les différents échantillons tout en mettant en avant les variables explicatives impactant le plus l'estimation.

Modèle	R^2	MAE	RMSE
Arbre de décision avec transformation de y	0.821	401.79	1066.07
Arbre de décision sans transformation de y	0.801	432.09	1121.69

TABLE 2 – Arbre de décision avec/sans transformation de y

Le tableau précédent nous amène à la conclusion qu'une transformation de la variable cible en informations par quantiles augmente légèrement les performances du modèle. Nous avons procédé, ensuite, à la recherche des meilleurs hyperparamètres conduisant à des valeurs de métriques plus faibles. Comme illustré dans le chapitre 3.3.1, une approche pourrait consister à employer la méthode GridSearchCV ou RandomSearchCV. Mais nous utilisons une version simplifiée de ces techniques par l'usage de la courbe de validation d'autant plus que le nombre d'hyperparamètres d'un arbre de décision n'est pas très important. Cette courbe montre la réponse des performances du modèle aux variations de la valeur d'un hyperparamètre. On commence par sélectionner un des hyperparamètres puis on représente la courbe de validation en fonction de celui-ci. Les deux hyperparamètres étudiés ici sont le nombre de feuilles maximum par noeuds *max_leaf_nodes* puis le nombre d'échantillons minimum dans une feuille *min_samples_leaf*. La figure 10 relève l'influence de la taille minimale d'échantillon par feuille dans la convergence du score alors que le nombre de noeud maximal ne semble pas influer sur la performance du modèle. Nous retenons ainsi une valeur de 10 pour l'hyperparamètre *min_samples_leaf*. Nous obtenons les performances suivantes :

Modèle	R^2	MAE	RMSE
Arbre de décision avec transformation de y, sans hyperparamétrage	0.821	401.79	1066.07
Arbre de décision avec transformation de y et *min_samples_leaf* $= 10$	0.875	305.11	888.05

TABLE 3 – Arbres de décision avec hyperparamétrage

7. https ://scikit-learn.org/stable/modules/generated/sklearn.compose.TransformedTargetRegressor.html

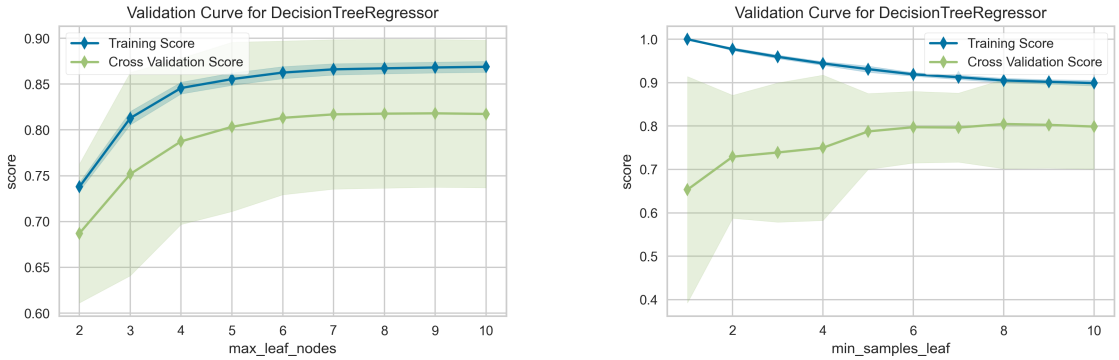


FIGURE 10 – Courbe de validation en fonction des hyperparamètres *max_leaf_nodes* et *min_samples_leaf*

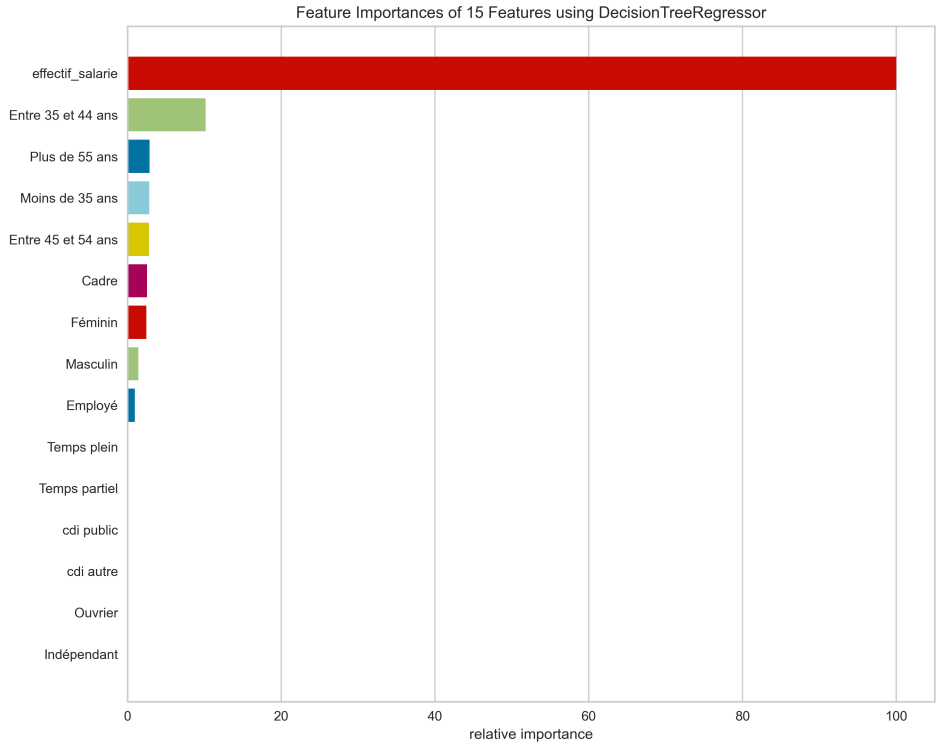


FIGURE 11 – Importance des caractéristiques pour l'arbre de décision (normalisation à 100)

Ce modèle affiche de bien meilleures performances que les modèles précédents. Pour mieux comprendre son fonctionnement, l'importance des caractéristiques est représentée dans la figure 11.

Pour rappel, cette mesure permet de déterminer les variables les plus contributives par degré d'importance dans la prédiction du modèle. La variable renseignant l'effectif des salariés dans une entreprise est la plus déterminante pour la prédiction du nombre d'arrêts au vu de ce graphique. En même temps, ce constat est assez logique puisqu'une entreprise de grande taille (plus de 1000 salariés) a nécessairement plus de jours arrêtés qu'une petite entreprise (notamment par l'effet taille). Par contre, en dessous de cette variable, deux conclusions intéressantes peuvent être tirées :

De l'une,

— le premier critère de division de l'arbre de décision est la proportion de personnes dans les différentes tranches d'âges,

— le second critère est manifestement la catégorie socio-professionnelle puis enfin vient le sexe. La catégorie cadre semble plus prédominante dans la prédiction des arrêts ensuite vient le sexe féminin.

De l'autre, le type d'emploi c'est-à-dire un contrat à durée indéterminée dans le public ou d'autres secteurs que le public/le privé n'est pas important. Le fait que l'emploi soit à temps partiel ou à temps plein ne prédomine cependant pas dans la prédiction du nombre annuel de jours d'absence par entreprise. La figure 12 donne la visualisation de cet arbre à une profondeur de deux feuilles maximum ($max\_depth = 2$) et d'un minimum de 10 observations pour chaque noeud. La prédiction à chaque noeud peut être observé par l'attribut value, présent dans chaque noeud, qui n'est rien d'autre qu'une moyenne des prédictions de l'ensemble de l'échantillon du noeud.

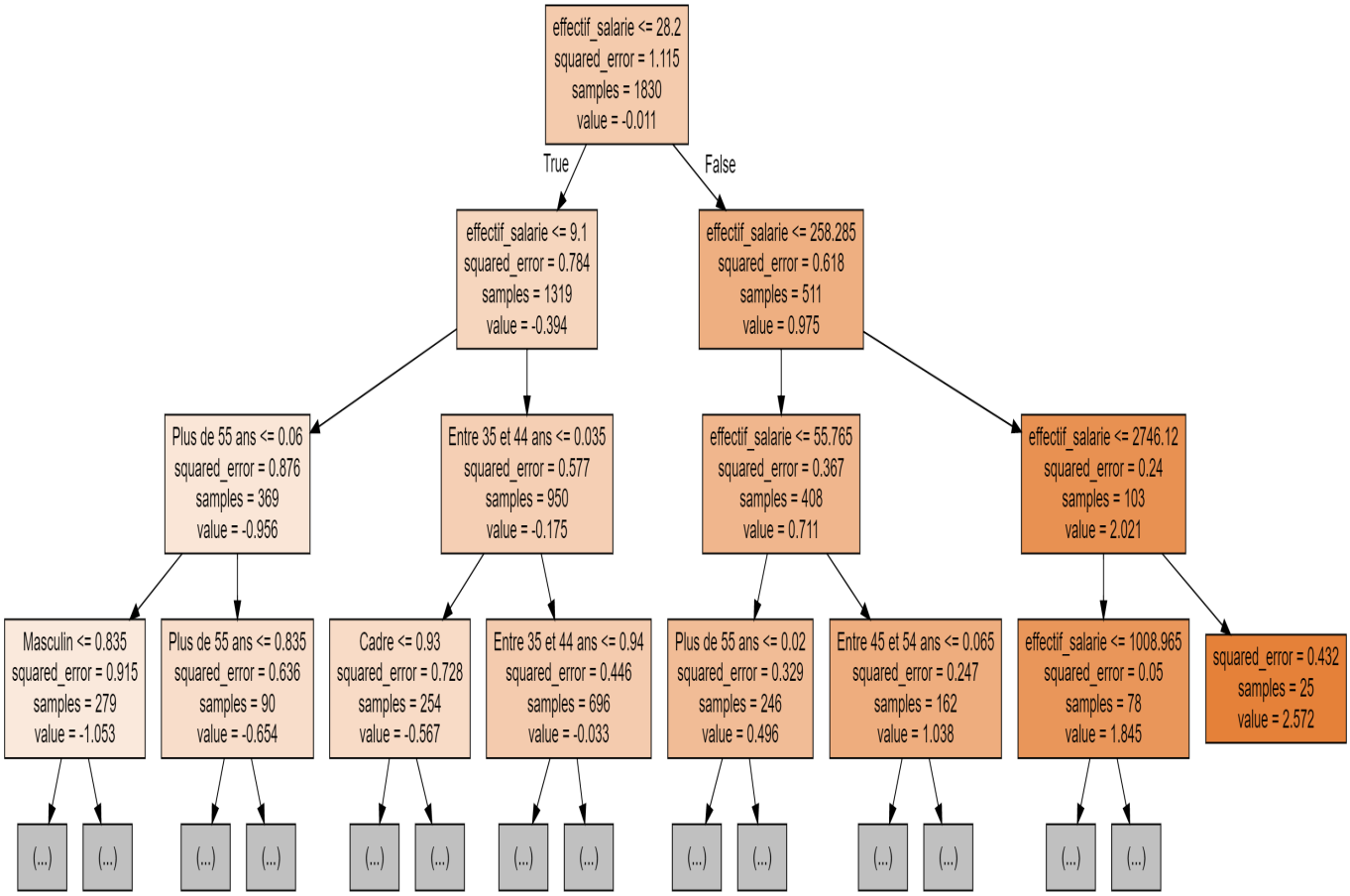


FIGURE 12 – Arbre de décision avec une profondeur maximale de 2, $max\_depth = 2$

La limite de la profondeur fixée à 2 permet de zoomer sur les caractéristiques les plus importantes comme elles sont souvent vers le haut de l'arbre. Hormis l'effectif salarié, on remarque une surreprésentation de la tranche d'âge. Le sexe ne vient qu'en deuxième position. Pour trouver les vraies valeurs prédites par le modèle, il faut traverser l'arbre depuis la racine jusqu'à la dernière feuille en suivant les critères mentionnés dans les noeuds.

La principale limite des CART est que sa structure n'est pas robuste. En effet, si on modifie une variable, et en particulier la variable qui se trouve au niveau du nœud racine, c'est toute la structure de l'arbre qui est modifiée. Plus que la prédiction, l'objectif premier de la modélisation par les CART est d'identifier les variables les plus discriminantes du jeu de données et d'isoler des groupes d'entreprises suivant leur comportement face au nombre

annuel de jours d'absence. Les résultats fournis par l'arbre de régression sont jugés insatisfaisants. Un seul arbre est utilisé pour obtenir ces résultats, l'agrégation par boosting doit permettre de les améliorer.

4.3.2 Ensemble learning de type Boosting

Le boosting est une autre technique d'agrégation de modèles. Elle présente l'intérêt qu'elle part de modèle très faible en général au début de l'apprentissage et parvient à rendre le modèle plus fort au fur et à mesure de l'entraînement. Un premier niveau d'analyse a été de juger des performances des différents algorithmes par rapport à trois métriques différentes : RMSE, le MAE puis le R^2.

Modèles	R^2	MAE	RMSE
Gradient Boosted Regression	0.919	279.85	716.49
Light GBM Regression	0.891	307.07	831.50
XGBoost Regression	0.903	300.66	782.08

Transformation **non appliquée** sur y

Modèles	R^2	MAE	RMSE
Gradient Boosted Regression	0.904	278.09	779.34
Light GBM Regression	0.890	283.59	834.61
XGBoost Regression	0.875	288.60	890.02

Transformation appliquée sur y

TABLE 4 – Modèles boosting sans hyperparamétrage

Le tableau permet de relever une nette amélioration des performances globales par rapport à un unique arbre de décision. On remarque également que les différentes variantes du gradient boosting affichent des valeurs en dessous de 300 unités pour la métrique MAE. Ce qui indique notamment une robustesse face aux valeurs aberrantes. Néanmoins une analyse comparative des deux tableaux suggère une détérioration du pouvoir explicatif du Gradient Boosting lorsque nous opérons une transformation sur la variable cible avant l'entraînement.

Dans les différents modèles illustrés ci-dessus, aucun réglage d'hyperparamètres n'a été effectué. C'est pour cela que nous avons procédé dans une seconde partie à la phase de recherche de meilleurs hyperparamètres. Pour pouvoir trouver les bons paramètres par la méthode GridSearchCV, un choix de valeurs a été établi sur cinq hyperparamètres que sont :

- n_estimators : c'est le nombre maximal d'arbres pour l'algorithme Boosting
- max_depth : la profondeur maximale de chaque arbre
- learning_rate : le taux d'apprentissage permettant de jouer sur la vitesse de convergence de l'algorithme
- num_leaves : le nombre de feuille maximal de l'arbre
- reg_lambda : Une régularisation de type L2 est appliquée sur le modèle pour favoriser les variables avec les petits poids dans le but de lutter contre le surapprentissage.

Face à la multitude de paramètres disponibles, nous nous sommes intéressés à quelques-uns et ils diffèrent d'un algorithme à l'autre.

- ✱ Pour le Gradient Boosting, nous avions considéré les valeurs suivantes :
  - 'n_estimators' : {100, 80, 60, 55, 51, 45}
  - 'max_depth' : {7, 8}
  - 'reg_lambda' :{0.26, 0.25, 0.2}
- ✱ Pour le XGBoost
  - 'n_estimators' : {100, 80, 60, 50, 45},
  - 'max_depth' : {3, 4, 5, 6},
  - 'learning_rate' :{0.2, 0.15}
- ✱ Pour le Light GBM
  - 'n_estimators' ∈ [30, 100],

— 'max_depth' : {-1, 3, 4},

— 'num_leaves' : {10, 11, 12},

— 'learning_rate' : {0.1, 0.01, 0.05}

Cette recherche par grille nous conduit au tableau de gauche présenté dans la table 5. Le tableau de droite correspond aux mêmes modèles sans hyperparamétrage.

Modèle	R^2	MAE	RMSE
Gradient Boosted Regression	0.921	269.917	709.293
Light GBM Regression	0.884	284.582	858.406
XGBoost Regression	0.890	278.035	834.271

Transformation **appliquée** sur y

Modèle	R^2	MAE	RMSE
Gradient Boosted Regression	0.902	284.19	788.71
Light GBM Regression	0.890	283.59	834.61
XGBoost Regression	0.875	288.60	890.02

Transformation appliquée sur y

TABLE 5 – Modèles boosting avec hyperparamétrage suivant le type de modèle par GridSearchCV

Mis à part le modèle LightGBM les performances des modèles ont été améliorées par les méthodes d'ensemble. La baisse des différents métriques du Light GBM peut notamment s'expliquer par du surapprentissage car en effet le modèle sans hyperparamétrage affiche une valeur plus élevée pour beaucoup de paramètres (*max_depth = -1*, ie infinie, *n_estimators = 100*, *num_leaves = 31*) et le fait de le confronter à la validation croisée couplée permet de relativiser sur les scores produits par le modèle. Si nous n'appliquons aucune transformation sur y nous obtenons le tableau suivant :

Modèle	R^2	MAE	RMSE
Gradient Boosted Regression	0.907	302.287	769.077
Light GBM Regression	0.892	294.943	828.265
XGBoost Regression	0.884	314.028	858.514

TABLE 6 – Modèles boosting avec hyperparamétrage sans transformation de y

A la différence des autres modèles, le light GBM se voit améliorer en termes de performance et affiche une RMSE plus faible. Ce modèle pénalise les erreurs les plus importantes et accorde une valeur particulière au RMSE. Ainsi, nous conservons les deux modèles et essayons d'évaluer leurs performances de façon plus globale et nous tenterons en dernier ressort de les tester sur l'ensemble de validation ("base que les deux n'ont guère venu jusqu'à présent"). Le modèle retenu est le suivant :

— GBM Regression avec y transformée, Les hyperparamètres qui ont contribués à cette performance sont :
- learning_rate = 0.15,
- max_depth = 3,
- n_estimators = 45

— Light GBM sans transformation de y Les hyperparamètres qui ont contribués à cette performance sont :
- n_estimators = 50,
- max_depth=3,
- num_leaves=10,
- learning_rate= 0.1

4.3.3 Évaluation des régressions

Dans la section précédente, il a été présenté les résultats des différents modèles. Un premier degré de comparaison a été fait à base des métriques tels que le RMSE, le MAE et le Score/R^2. Un deuxième niveau d'analyse peut être fait en observant le graphique des résidus. Ce diagramme a l'intérêt de représenter la distribution des erreurs autour de

l'origine. Elle permet de déceler les zones de concentration des erreurs et de montrer les erreurs aberrantes susceptibles d'avoir un impact important dans le modèle.

Le diagramme des résidus pour le GBM est le suivant :

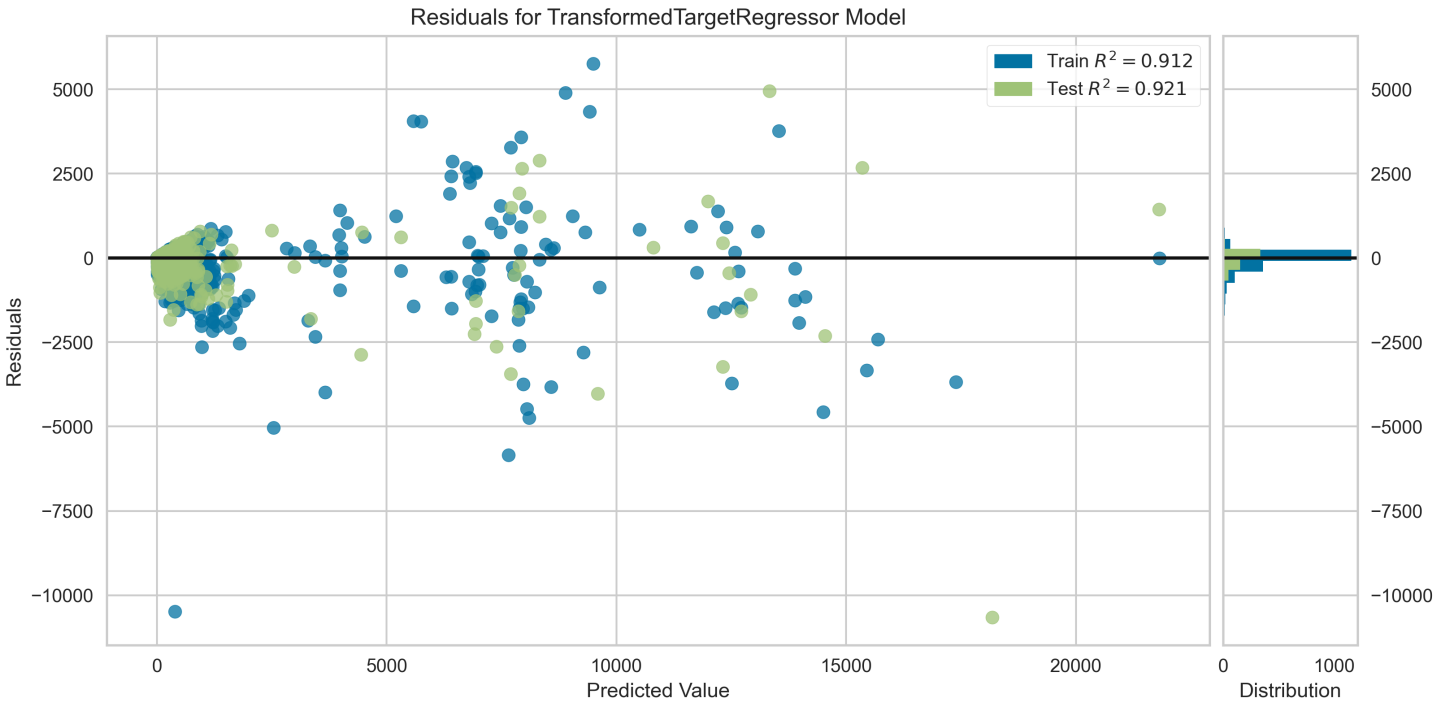


FIGURE 13 – Diagramme des résidus pour le Gradient Boosting Machine

Les erreurs ne sont pas réparties au hasard néanmoins on observe une concentration des prédictions importantes de moins de 1000 arrêts. Ce qui peut être expliquée par la structure de la base de données. Néanmoins l'influence que peut avoir les valeurs aberrantes dans la prédiction est atténuée. Ce modèle sera en effet plus robuste pour la prédiction des valeurs abérrantes. Une deuxième conclusion que l'on peut tirer de ce modèle est qu'un modèle linéaire ne semble pas adapté à la modélisation du nombre d'arrêts. Autrement dit, il est moins probable qu'il subsiste une forte relation linéaire entre le nombre annuel de jours d'absence et les déterminants socio-économiques. Si l'on s'intéresse à l'importance des variables en pourcentage de la caractéristique principale, on obtient le graphe 14.

On remarque de ce graphe que outre les variables telles que l'effectif salarie et les différentes tranches d'âge qui apparaissaient dans le graphique 11, des nouvelles variables font figure tels que le temps consacré à l'emploi (temps plein ou temps partiel). On remarque également la présence de la catégorie Ouvrier qui est également présente mais relativement peu importantes. De façon globale nous retenons que la tranche d'âge est la plus importante des variables, à part l'effectif salarie, ensuite le sexe est également la deuxième variable déterminante puis vient la catégorie socio-professionnelle. Ce n'est qu'après on retrouve le temps consacré à l'emploi (temps plein/temps partiel) et enfin on retrouve la csp Ouvrier. Néanmoins le sens des impacts n'est pas connu, pour cela, on peut avoir recours à la librairie Shap tel qu'explicité dans la section 3.4.2.

Explicabilité du modèle de GBM par la librairie SHAP

De façon résumée, l'approche shapley privilégie la compréhension d'un modèle au détriment d'une comparaison de modèles. Elle permet d'expliquer les prédictions individuelles tout en procurant une vision globale du modèle. nous obtenons le diagramme 15.

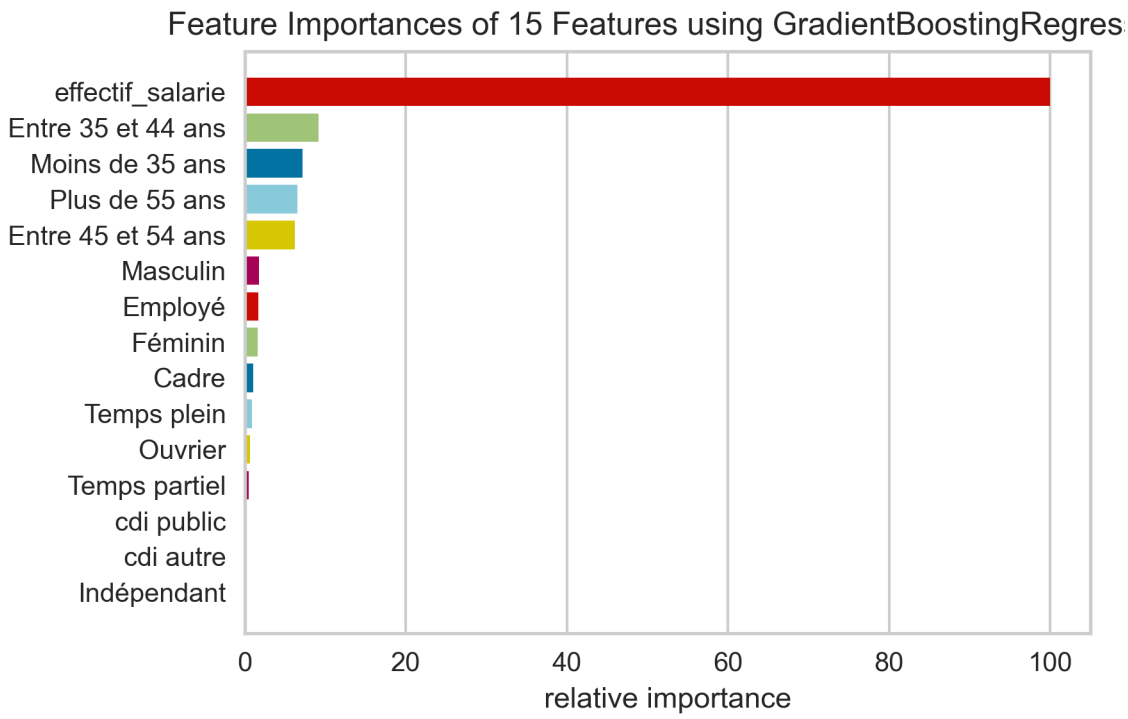


FIGURE 14 – Importance des caractéristiques obtenue par le GBM

Le graphique récapitulatif combine l'importance des caractéristiques et leurs effets. Chaque point du graphe récapitulatif est une valeur de Shapley pour une caractéristique et un échantillon de la base d'apprentissage. La position sur l'axe des y est déterminée par la variable et sur l'axe des x par la valeur de Shapley. La couleur représente la valeur de la caractéristique, de faible à élevée. Les points qui se chevauchent sont répartis dans la direction de l'axe des ordonnées, ce qui nous donne une idée de la distribution des valeurs de Shapley par variable. Les variables sont ordonnées en fonction de leur importance.

Le graphique nous permet de conclure à notre première hypothèse sur l'impact positif de l'effectif des salariés d'une entreprise. Les entreprises affichant de fortes proportions pour la catégorie de plus de 55 ans ont un impact positif sur le nombre annuel de jours d'absence. Par contre la proportion des salariés ayant un âge compris entre 35 et 44 ans fait plutôt baisser ce nombre notamment puisque nous observons un léger étalement vers la droite avec une accentuation de la couleur rouge. La proportion de cadre dans une entreprise est un indicateur de baisse du nombre annuel de jours d'absence. La proportion de personnes de sexe masculin fait baisser ce taux tandis que la proportion de sexe masculin fait augmenter ce taux et son effet est beaucoup plus important que celui de sexe féminin. Pour les variables telles que la proportion d'ouvriers, la proportion de salariés à temps plein, il n'est pas aisé du moins graphiquement de déceler la nature de la corrélation. Une alternative est de construire le diagramme de dépendance SHAP pour ces variables. Il montre l'effet marginal d'une ou deux caractéristiques sur le résultat prédit d'un modèle d'apprentissage automatique. Il indique également si la relation entre la cible et une caractéristique est linéaire, monotone ou plus complexe.

Dans le graphique 16, nous représentons le diagramme de dépendance pour chacune de ces deux variables.

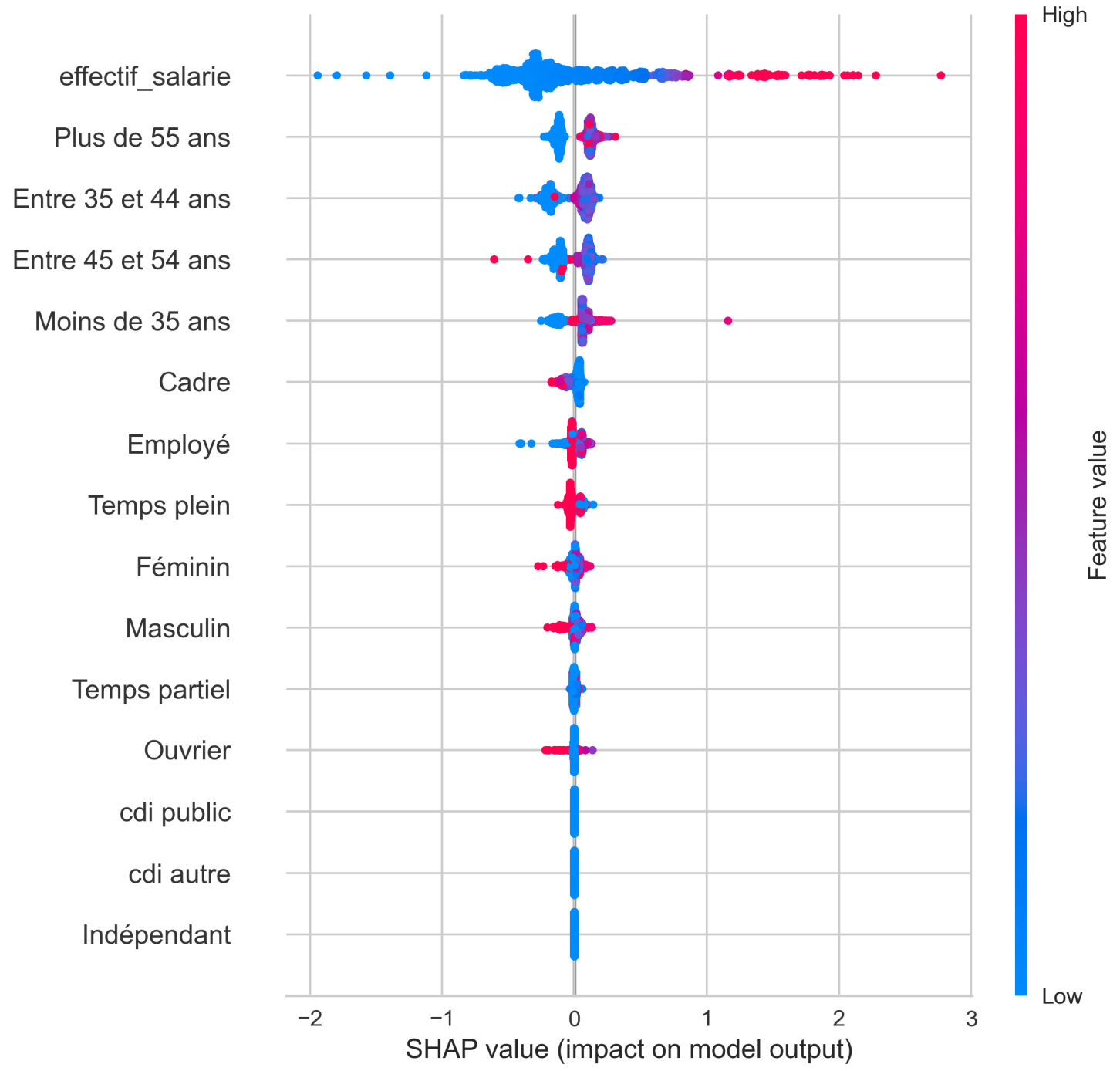


FIGURE 15 – Diagramme de synthèse SHAP

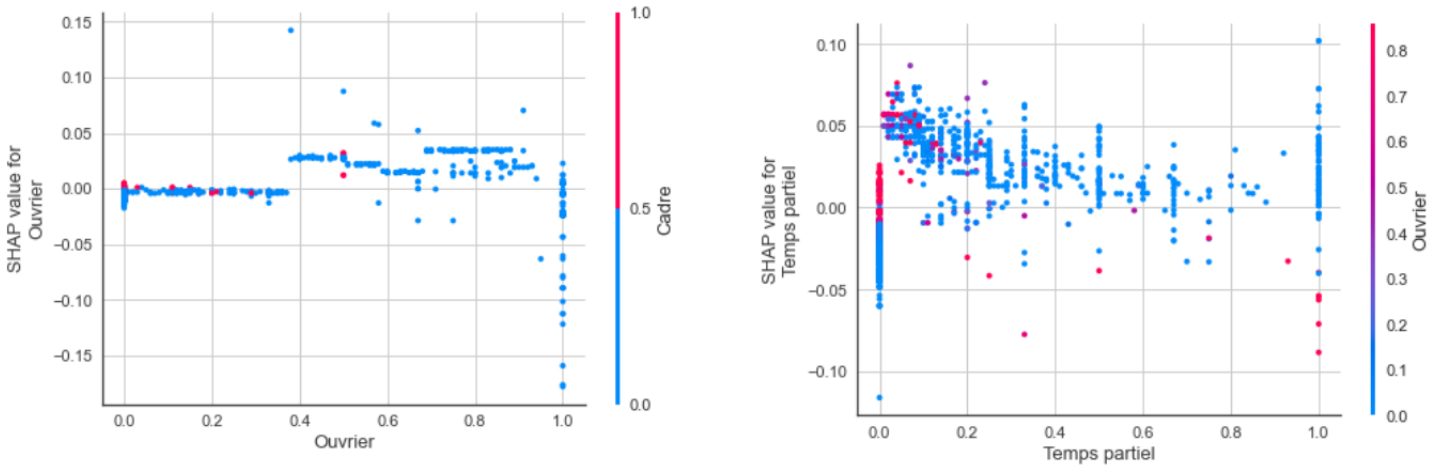


FIGURE 16 – Diagramme de dépendance des variables Temps partiel et Ouvrier

L'intérêt de ce diagramme est qu'en plus de mettre en exergue la relation existante entre la variable d'intérêt – le temps partiel/la csp Ouvrier – et la cible – le nombre d'arrêts annuel par entreprise – il inclut automatiquement une autre variable avec laquelle la variable choisie interagit le plus. Le graphique ci-dessus ne confirme qu'une complexité dans la relation entre la variable proportion d'Ouvrier et le nombre annuel de jours d'absence. En revanche, lorsque la proportion de salariés employées à temps partiel augmente, on note une baisse progressive de la valeurs des Shap. Cela indique notamment une baisse aussi de la valeur du nombre annuel de jours d'absence. Une explication plausible serait de se dire que lorsqu'on est employé en temps partiel, on est plus enclin à prétendre moins aux arrêts maladie puisque forcément on est moins payé (car à temps partiel comparativement aux employés à temps plein).

En résumé, nous proposons le tableau suivant qui récapitule l'ensemble des caractéristiques importantes dans la prédiction du modèle Gradient Boosting ainsi que la nature de leur influence dans la prédiction ou plus précisément la nature de leur corrélation avec la variable d'intérêt qu'est le nombre d'arrêts annuel.

Variable	Corrélation
Effectif salarié	+++
Plus de 55 ans	++
Cadre	- -
Masculin	- -
Féminin	-
Temps partiel	-

Note : Le nombre de +/- correspond au degré de la corrélation

TABLE 7 – Récapitulatif de l'influence des variables sur la variable à expliquer

Nous reproduisons la même analyse avec le modèle Light GBM sans transformation de la variable cible. Le diagramme des résidus obtenus pour ce modèle est présenté dans la figure 17. La variable cible n'étant pas modifiée avant l'application du modèle, les zones de concentration des erreurs se distribuent légèrement sur toute l'étendue de la variable cible. Néanmoins la concentration des erreurs est toujours accrue en dessous de 1000 arrêts annuels. Hormis cette zone, la distribution des erreurs paraît aléatoire.

L'importance des caractéristiques est visualisée sur le graphique 18. Dans l'ordre d'importance, nous avons l'effectif salarié, la proportion des personnes ayant plus de 55 ans, la proportion des personnes à temps partiel, les proportions de salariés cadre, ouvrier et employé, cadre et Ouvrier avant que s'en suivent le sexe Féminin puis masculin. Ainsi les résultats présentés dans ce modèle diffèrent significativement de ceux obtenus à partir du modèle GBM. On peut noter également le même dégré d'importance pour la proportion de salariés cadre et employé. Le graphique de Shapley

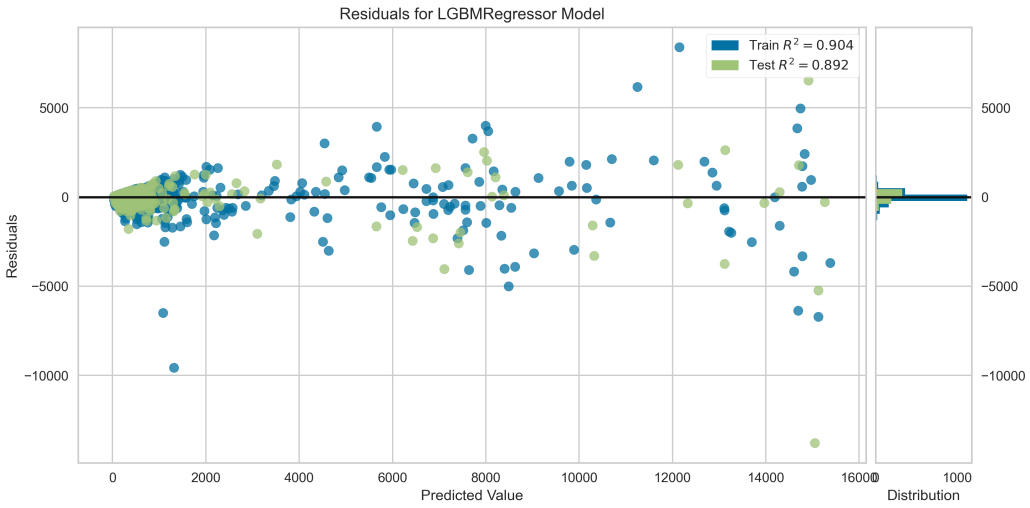


FIGURE 17 – Graphique des résidus

obtenu pour ce modèle s'avère difficilement interprétable pour l'ensemble des variables explicatives autre que l'effectif salarié.

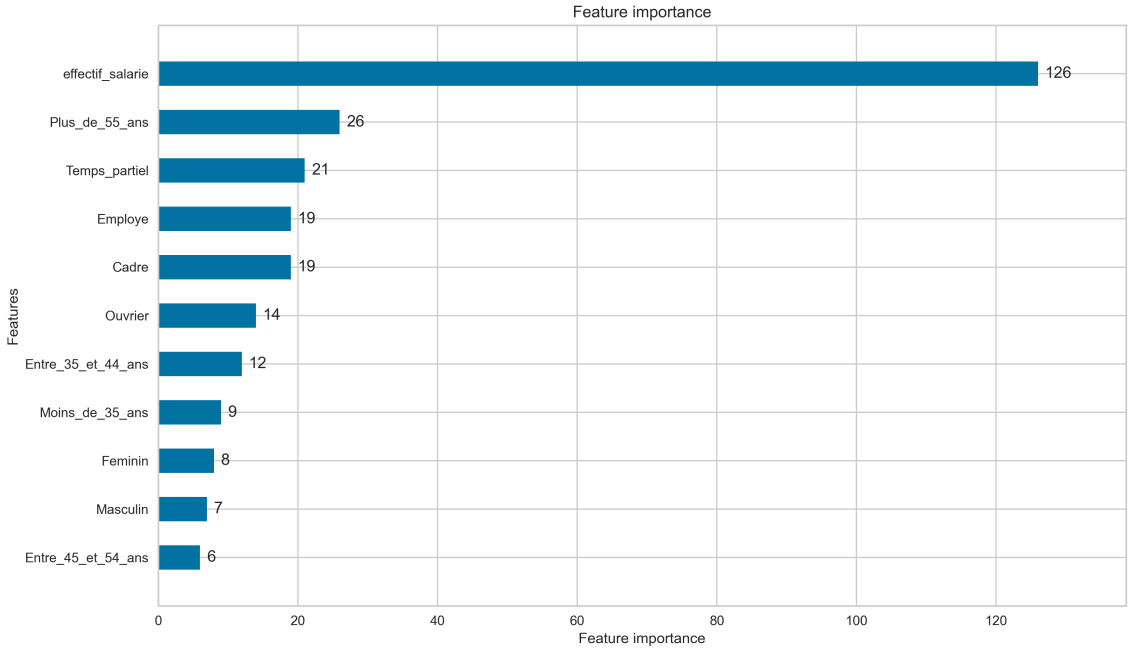


FIGURE 18 – Importance des caractéristiques du LightGBM

Nous avons évalué ces deux modèles sur l'ensemble de test qui représente 20% des données que le modèle n'a guère vu jusqu'à présent. Les performances affichées sont de 86% pour le R^2 du Gradient Boosting et de 85% pour le Light GBM. En définitive, ce sera le modèle de type boosting Gradient Boosting qui sera retenu pour ces performances et

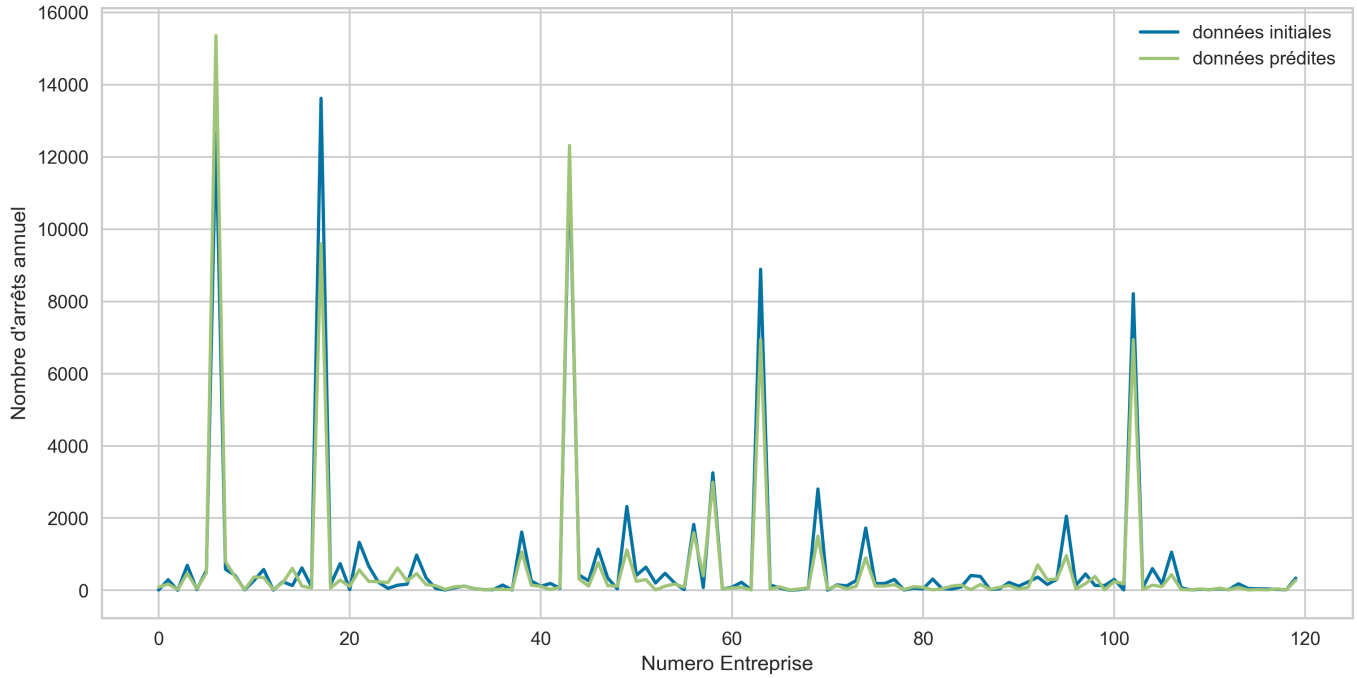


FIGURE 19 – Trajectoires du nombre annuel de jours d'absence initial

ses interprétations. Le graphe 19 montre l'évolution de deux courbes dont l'une est fonction du nombre annuel de jours d'absence pour les 120 premières entreprises et l'autre qui est fonction du nombre d'arrêts annuel prédits par le modèle Gradient Boosting. La conclusion qu'on pourra tirer de ce modèle est qu'il capte assez bien la dynamique derrière le nombre annuel de jours d'absence.

Conclusion

En somme, ce travail a permis de mettre en relief, les différents déterminants dans l'explication des arrêts maladie. Il a permis de relever l'influence des déterminants tels que la tranche d'âge et la catégorie socio-professionnelle sur le nombre de jours d'absence annuel mais de confirmer également la pertinence de la variable telle que le temps consacré à l'emploi. Les différentes variables comme le contrat de travail à durée indéterminée dans le privé ou dans d'autres secteurs, autre que le public, n'influent pas significativement sur la variable d'intérêt. Les techniques d'interprétation de modèles ont contribué à l'explication de ces modèles de machine learning ainsi que le sens de leur influence.

Le diagnostic des différents modèles par le diagramme des résidus, pour le modèle retenu, soutient une relation potentiellement non linéaire entre les caractéristiques considérées et le nombre d'arrêts annuel. Il souligne également une part inexpliquée importante dans les résidus. De cette conclusion, nous pourrions tirer que les variables explicatives sont certes représentatives mais demeurent insuffisantes pour expliquer les déterminants de l'arrêt. Pour pallier ce point, il pourrait être plus judicieux d'ajouter des variables tels que le secteur d'activité de l'entreprise, de proposer des variables mesurant le degré de satisfaction des salariés au sein d'une entreprise, la santé mentale du salarié, etc. Ce mémoire n'a pas abordé ces aspects par manque d'accessibilité à ces informations dans les bases de données DSN mises à disposition pour l'étude. Un autre point serait également de prendre en compte dans la modélisation les différentes interactions entre les variables explicatives. Il peut s'agir par exemple de voir l'influence d'être homme à temps plein/partiel ou encore d'être femme cadre/non cadre dans la prédiction de l'arrêt. L'interprétation de ces interactions dans des modèles de machine learning reste aujourd'hui complexe comme le souligne Kshitij et al.(2007). Ceci justifie le fait que l'on ait pas eu recours à ce type d'analyse dans les travaux présentés dans ce mémoire.

Enfin, cette étude nous permet de faire une recommandation qui est celle de regrouper en différentes classes les variables telles que l'âge, la catégorie socioprofessionnelle afin d'améliorer la prédiction des modèles.

Bibliographie

[1] E. Demou, S. Smith et al. (2018), *Evaluating sickness absence duration by musculoskeletal and mental health issues : a retrospective cohort study of Scottish healthcare workers*, BMJ open, vol. 8, no. 1, p. e018085.

[2] M. Virtanen, J. Ervasti et al., (2018), *Lifestyle factors and risk of sickness absence from work : a multicohort study*, The Lancet. Public Health, vol. 3, no. 11, p. e545–e554, 2018.

[3] W. Beemsterboer, R. Stewart et al. (2009), *A literature review on sick leave determinants (1984-2004)*, International Journal of Occupational Medicine and Environmental Health, vol. 22, no. 2, p. 169–179, 2009.

[4] H. de Vries, A. Fishta et al. (2018), *Determinants of Sickness Absence and Return to Work Among Employees with Common Mental Disorders : A Scoping Review*, Journal of Occupational Rehabilitation, vol. 28, no. 3, p. 393–417, 2018.

[5] J. Petersen, L. Kirkeskov, et al.(2019), *Physical demand at work and sick leave due to low back pain : a cross-sectional study*, BMJ open, vol. 9, no. 5, p. e026917.

[6] A. Alipour, M. Ghaffari, B. Shariati, I. Jensen et E. Vingard, *Four-year incidence of sick leave because of neck and shoulder pain and its association with work and lifestyle*, Spine, vol. 34, no. 4, p. 413–418, février 2009.

[7] J. Bué, T. Coutrot, N. Guignon et N. Sandret, *Les facteurs de risques psychosociaux au travail*, Revue française des affaires sociales, no. 2, p. 45– 70, 2008, publisher : La Documentation fran¸caise. [En ligne]. Disponible : https://www.cairn.info/revue-francaise-des-affaires-sociales-2008-2-page-45.htm

[8] F. W. O'Reilly et A. B. Stevens, *Sickness absence due to influenza* Occupational Medicine (Oxford, England), vol. 52, no . 5, p. 265–269, août 2002.

[9] S. Thewissen, D. MacDonald, C. Prinz et M. Stricot, *The critical role of paid sick leave in the COVID-19 health and labour market crisis*, juill. 2020. [En ligne]. Disponible : https://voxeu.org/article/paid-sick-leave-during-covid-19-health-and-labour-market-crisis

[10] Aurélien Géron. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow, Published by *O'Reilly Media*, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[11] C. Molnar, Interpretable Machine Learning, 2019, (consultée le 31 octobre 2021), [En ligne], *chapitre : Local Model-Agnostic Methods*. Disponible : https://christophm.github.io/interpretable-ml-book/

[12] Harisson Matt. (2019). Machine learning : Les fondamentaux - Exploiter des données structurées en python, Mars 2019 Publisher(s) :*Editions First*, ISBN : 9782412056028

[13] V. Čike, H. Maškarin Ribarić et K. Črnjar, 2018, *The Determinants and Outcomes of Absence Behavior : A Systematic Literature Review*, Social Sciences, vol. 7, no . 8, p. 120, number : 8 Publisher : Multidisciplinary Digital Publishing Institute. [En ligne]. Disponible : https://www.mdpi.com/2076-0760/7/8/120

[14] A. C. Group, "7 ème baromètre de l'Absentéisme," 2015. [En ligne]. Disponible : http://presse.ayming.com/mobilerelease.aspx?ID=34695

[15] Shapley, L. S.. "17. A Value for n-Person Games". *Contributions to the Theory of Games (AM-28), Volume II*, edited by Harold William Kuhn and Albert William Tucker, Princeton : Princeton University Press, 1953, pp. 307-318. https://doi.org/10.1515/9781400881970-018

[16] Fernando López, (page consultée le 31 octobre 2022), SHAP : Shapley Additive Explanations, [en ligne] https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

[17] Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

[18] Ribeiro, M. T., Singh, S., Guestrin, C. (2016, August). " Why should i trust you ?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

[19] Duchemin T., 2010., *Méthodologie d'analyse et de surveillance pour la prévention des arrêts maladie.* Ph.D. Dissertation, École doctorale Sciences et Méiers de l'Ingénieur, Laboratoire MESuRS, CNAM.

[20] Loureiro D., 2021, *Utilisation de la DSN et de l'open data pour élaborer et expliquer un zonier incapacité.* Mémoire d'actuariat, ENSAE Paris - IP Paris, chapitre 2 & 3.

[21] Andre S., 2019, *Comportement de versement libre en épargne individuelle : approche conceptuelle et modélisation.* Mémoire d'actuariat, ISFA Lyon.

[22] Duchemin T. , A. Bar-Hen et al., (2019), *Hierarchizing Determinants of Sick Leave : Insights From a Survey on Health and Well-being at the Workplace*, Journal of Occupational and Environmental Medicine, vol. 61, no . 8, p. e340–e347.

[23] Alvarez-Melis, David, and Tommi S. Jaakkola. (2018). *On the robustness of interpretability methods*, arXiv preprint arXiv :1806.08049.

[24] Sophie Hilgard, Emily Jia et al.(2020). *Fooling lime and shap : Adversarial attacks on post hoc explanation methods.* In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186.