Titre: Title:	Outil de diagnostic et de prévention de l'insatisfaction des employés
Auteur: Author:	Guillaume Vergnolle
Date:	2020
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Vergnolle, G. (2020). Outil de diagnostic et de prévention de l'insatisfaction des employés [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/5349/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/5349/
Directeurs de recherche: Advisors:	Nadia Lahrichi
Programme: Program:	Maîtrise recherche en mathématiques appliquées

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Outil de diagnostic et de prévention de l'insatisfaction des employés

GUILLAUME VERGNOLLE
Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Mathématiques appliquées

Août 2020

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Outil de diagnostic et de prévention de l'insatisfaction des employés

présenté par **Guillaume VERGNOLLE**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Louis-Martin ROUSSEAU, président
Nadia LAHRICHI, membre et directrice de recherche
Philippe LANGLAIS, membre

DÉDICACE

À mon rayon de soleil, nos rires, nos voyages,
nos souvenirs et nos chocolats chauds

REMERCIEMENTS

Je souhaite commencer par remercier la professeure Nadia Lahrichi pour m'avoir encadré tout au long de ce projet, dont les commentaires et le suivi ont été constructifs et formateurs afin de me faire avancer.

Merci à toute l'équipe d'AlayaCare de m'avoir accueilli chez eux et de m'avoir formé : Simon Germain pour ton accueil et tes conseils, Pierre Leca pour m'avoir accordé beaucoup de temps au milieu d'un emploi du temps déjà bien rempli, Jonathan Vallée pour ton soutien, Marc-André Geraldo véritable camarade de sciences de données et ami, Adam Cherti pour ton analyse fine et tes idées de génie, Jie-Chen Wu pour le recul que tu m'as donné sur mon étude à travers la tienne, Naomi Goldapple pour ton énergie et tes encouragements, Simon Maxwell pour m'avoir aiguillé sur certaines méthodes et répondu à mes questions, et à tous les autres employés que j'ai côtoyé autour d'un jeu de plateau ou d'une bière.

Enfin, merci à ma famille, mes amis, et tous ceux qui m'ont entouré pendant ce projet. Merci à toutes les personnes qui m'ont aidé directement ou indirectement, c'est aussi grâce à vous que j'ai pu écrire ce mémoire.

RÉSUMÉ

La rétention d'employés dans le domaine de la santé, notamment dans les centres de soins à domicile, est un des problèmes majeurs du secteur depuis la fin du XXe siècle. De nombreuses études se sont déjà penchées sur ce problème afin de mieux comprendre ce phénomène. Les facteurs identifiés peuvent être regroupés en 5 catégories :

— les caractéristiques de l'employé
— la planification des visites et des horaires
— les caractéristiques du métier
— les interactions sociales inter-professionnelles et avec les patients
— l'impact de la structure dans laquelle l'employé travaille

Ces études tentent de qualifier la satisfaction des employés dans leur emploi à travers des questions adressées sous forme de sondages. Les réponses sont ensuite analysées avec des méthodes statistiques afin de déterminer les causes du problème de la rétention d'employés.

Cette étude propose une nouvelle approche du départ des employés en soins à domicile en utilisant directement leurs données d'activité : heures de travail, patients vus, disponibilités... Notre partenaire industriel AlayaCare est une compagnie montréalaise qui collecte depuis 2014 l'activité de ses clients : des centres de soins à domicile au Canada, aux États-Unis et en Australie. Les utilisateurs principaux de la plateforme sont les gestionnaires ou les coordonnateurs. Afin de répondre aux besoins en visites des patients, ils peuvent assigner des visites à des employés et ainsi avoir un outil efficace pour gérer les emplois du temps de chacun. En exploitant diverses informations comme la quantité de visites à chaque période, les distances parcourues, la variété des tâches réalisées ou encore les plages horaires travaillées, l'objectif est d'extraire les variables qui ont un impact sur la satisfaction des employés.

La modélisation proposée est un problème de classification à deux classes. Un des défis de ce projet est qu'aucune mesure évidente de l'intention de départ d'un employé au cours du temps n'est disponible. La solution envisagée est de distinguer les périodes qui ont précédé le départ d'un employé de celles plus anciennes. Ainsi, l'objectif est de relever et d'identifier les variables qui ont évolué au fil des périodes et qui pourraient expliquer pourquoi la situation d'un employé s'est détériorée, amenant à un départ volontaire.

Une étape de pré-traitement et de mise en forme des données a d'abord été mise en place afin de créer des ensembles d'entraînement pour les modèles de classification. Différents découpages temporels ont été testés afin de déterminer celui qui est le plus adapté à notre modélisation. Les modèles utilisés pour répondre au problème de classification sont :

— la régression logistique
— modèles d'arbres (forêts aléatoires, XGBoost)
— les réseaux de neurones

L'objectif est de sélectionner le modèle le plus adapté au problème de rétention d'employés tel que nous l'avons défini et d'explorer des espaces de paramètres pour obtenir les meilleures performances possibles. Ces expériences sont effectuées pour 10 agences de soins à domicile différentes, dont 6 se situent au Canada, 2 aux États-Unis et 2 en Australie. Comparer les résultats à travers les différents clients permet de généraliser les observations effectuées.

La stratégie de sélection de modèle implémentée a montré que l'architecture XGBoost obtient généralement les meilleures performances. Le découpage qui apportait le plus de signal utilise les données des 24 dernières semaines d'un employé. Les solutions ne permettent pas de répondre parfaitement au problème modélisé, mais les résultats obtenus permettent d'apporter des éléments de réponse. La variable la plus significative est liée au nombre de visites sur une période qui doit être élevé : ce résultat s'est généralisé à travers les différents clients et semble capital. Intuitivement, cette variable est directement relié à la rémunération, un grand nombre de visites apporte plus de sécurité. Les autres conclusions dépendent généralement d'un client à l'autre, on relève les pratiques suivantes qui ont permis d'augmenter la satisfaction des employés de manière générale :

— varier les durées de visites
— avoir des visites plus longues
— communiquer ses indisponibilités aux gestionnaires chaque semaine

Les autres critères mesurés ont soit une importance faible dans la classification, soit les résultats varient d'une structure à l'autre et sont difficiles à généraliser. De cette manière, on est capable de faire un diagnostic personnalisé pour chaque agence de soins ayant suffisamment de données afin d'identifier les facteurs ayant le plus d'impact sur la satisfaction de ses employés.

Le dernier chapitre propose d'appliquer ces résultats dans un algorithme d'optimisation présent chez AlayaCare. Celui-ci permet de prendre en compte les besoins et disponibilités des patients et des employés sur un horizon de plusieurs jours afin de créer les emplois du temps qui minimisent une fonction de coût. Une nouvelle formulation du coût associé aux routes prenant en compte les facteurs qui influencent le départ des employés identifiés par notre méthode est proposée. En améliorant l'algorithme d'optimisation, on souhaite créer des emplois du temps attentifs aux besoins des patients et aux préférences des employés. On pourrait alors espérer voir le taux de rétention augmenter et une meilleure satisfaction globale à travers les employés afin d'aider un secteur qui a plus que jamais besoin de main-d'oeuvre.

ABSTRACT

Employee retention is a burning problem in home care industry. Several studies have already been looking for solutions. They identified factors that can be grouped into five categories :

— employee characteristics
— scheduling
— job related factors
— social interactions with other employees and patients
— structure related factors

These studies aim for a measure of satisfaction using polls and surveys. The answers are then analyzed to highlight the causes of employee churn.

Our project proposes a new way of studying employee satisfaction by looking directly at employees visit schedule. A partnership with AlayaCare allowed us to have access to such data. This company from Montreal offers a platform to home care agencies to ease coordinators' workload. The main user is generally a manager or a coordinator. Through the software, one can assign employees to visits and have a better overview of the staff workload. Various information can be found in the database : the length of visits, the location, or the type of service given. The objective is to extract as many features as possible to identify important ones among them.

We decided to model this problem as a classification problem. One of the biggest challenge is that no obvious measure is available to assess the satisfaction of the employees through time. The proposed solution is to try to distinguish periods before an employee left from previous ones. Therefore, we should be able to identify the variables that evolved through time and may explain why an employee left.

Preprocessing and shaping the data are necessary steps before creating training sets for classification models. Different time divisions has been tried to determine which one fits best the way we modelled this problem. Several classification algorithms have been implemented :

— logistic regression
— tree-based models (random forests and XGBoost)
— neural networks

Model selection run on different hyper-parameters in order to reach optimal performances. This method is applied on 10 health care clients : 6 located in Canada, 2 in the United Stated, and 2 in Australia. Comparing the results across these companies may help to generalize the

observations.

The XGBoost architecture give the best results over the experiments. It's also recommended to use the last 24 weeks of an employee to better measure the evolution of his/her satisfaction through time. The solutions found don't match perfect classification, but they lead to some conclusions. The number of visits over a period is the key : this observation generalized over all clients. Among others, the following statements has been identified to improve one's satisfaction but may slightly differ a client from another:

— having different visit lengths through a period
— having longer visits
— communicate one's unavailabilities for a week

The other factors either have low importance or are different among clients and thus don't generalize well. In this way, we are able to make a custom diagnosis for each company, identifying the factors that matter.

The last section is dedicated to put the previous results into an optimization algorithm available at AlayaCare. By taking into account availability and needs of patients and employees, it creates daily schedules over a fixed period. A new cost formulation for routes is proposed by adding the necessary variables identified in our study. Regular use of this algorithm could improve the staff satisfaction while meeting patient needs, we would expect the turnover rate to decrease over time. It is hoped that such a solution could help an industry that lacks workforce more than ever.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

AUC	*Area Under Curve*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
ENN	*Edited Nearest Neighbors*
HHCRSP	*Home Health Care Routing and Scheduling Problem*
IRSAD	*Index of Relative Socio-economic Advantage and Disadvantage*
MBI	*Maslach Burnout Inventory*
MLP	*Multi Layer Perceptron*
MMSS	*McCloskey/Mueller Satisfaction Scale*
NWI	*Nursing Working Index*
NWI-R	*Revisited Nursing Working Index*
PES-NWI	*Practice Environment Scale of the Nursing Working Index*
RTA	Région de Tri d'Acheminement
SMOTE	*Synthetic Minority Over-sampling Technique*
ROC	*Receiver Operating Characteristic*
XGBoost	*Extreme Gradient Boosting*

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

L'industrie des soins à domicile est en plein essor au Canada. Les patients considèrent leur domicile comme un lieu privilégié pour combattre des maladies de longue durée ou pour y terminer ses jours en comparaison avec le milieu hospitalier. Un des plus grands défis de cette industrie est le vieillissement de la population puisque 70% des soins à domicile sont administrés à des personnes âgées de 65 ans et plus [1]. Le nombre de postes vacants pour le personnel infirmier a augmenté de 77 % entre 2015 et 2019 : l'embauche stagne depuis plusieurs années [2]. En parallèle, sur l'année 2015, plus de 400000 canadiens n'ont pas été satisfaits des soins prodigués à domicile, ce qui représente une personne sur trois. La principale cause de cette insatisfaction est l'indisponibilité du personnel infirmier [3].

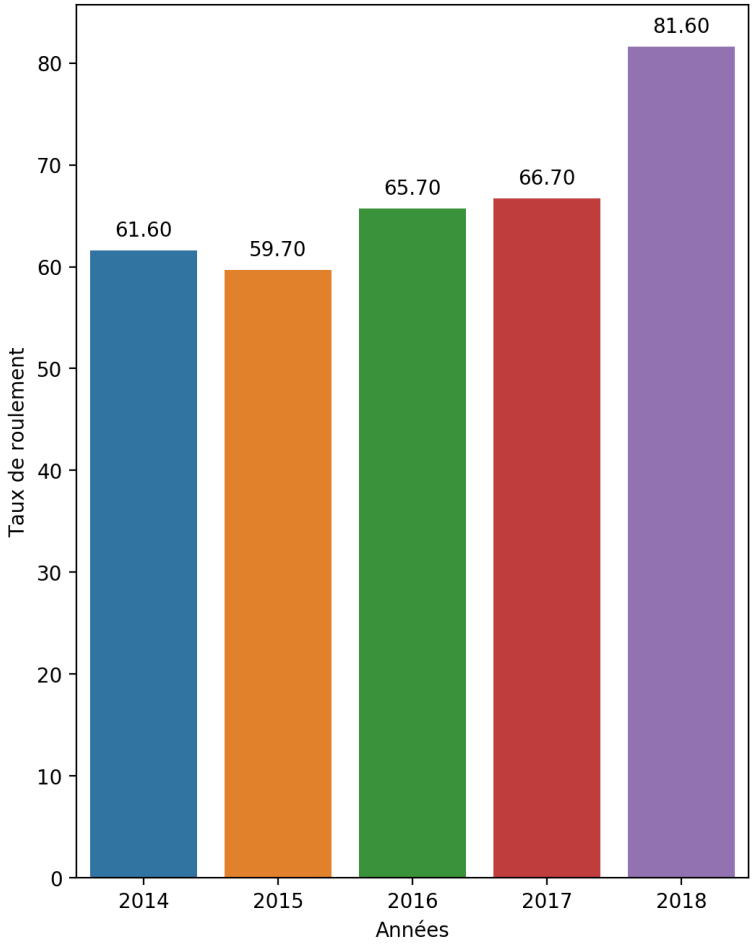

Figure 1.1 Évolution du taux de roulement en soins à domicile

Cette situation doit aussi faire face à une augmentation de l'insatisfaction du personnel soignant. La figure 1.1 montre l'évolution du taux de roulement de ces dernières années d'après les données de [4]. Le taux de roulement est calculé en divisant le nombre d'employés ayant quitté une agence par le nombre moyen d'employés sur une année. Face à une augmentation du taux de roulement dans les établissements de soins à domicile au Canada, les dirigeants des différents centres se focalisent sur les besoins et les préférences des employés afin de les conserver au sein de leur structure. Ceci passe notamment par l'aménagement des horaires de travail, une sélection attentive des jours de repos, ou encore une affectation régulière à certains patients. En effet, la perte d'un employé représente plusieurs coûts :

— formalités administratives pour terminer un contrat
— effort en ressources humaines pour rechercher de nouveaux candidats
— formalités administratives pour créer un nouveau contrat
— formation de l'employé à son nouveau poste
— rupture de la continuité des soins pour certains patients

Selon la taille de la compagnie, il peut être difficile de faire le suivi de chaque employé et de maîtriser l'état de satisfaction de sa force de travail. De nombreuses études ont déjà été menées afin de relever le niveau de satisfaction des employés au sein d'une structure à travers un questionnaire, mais le processus de récolte de données est long et ne peux pas être répété trop souvent dans le temps. Le développement d'un outil capable d'approximer l'état de satisfaction d'un employé à partir des données d'activité permettrait aux agences de soins à domicile d'adapter les assignations de visites afin de maximiser la satisfaction de sa force de travail. Réduire le nombre de départs des employés permettrait de garantir une meilleure continuité de soins auprès des patients et d'offrir une meilleure stabilité économique pour l'employé et la structure de soins.

La présente étude utilise les données de l'entreprise AlayaCare. Celle-ci propose une plateforme de gestion à des cliniques de santé à domicile afin de simplifier l'organisation de leur charge de travail et le suivi de leurs patients. Fondée en 2014, les données collectées chez leurs clients ont été récupérées à des fins de recherche. La base de données initiale comprend 12 millions de visites réparties parmi 30000 employés dans 117 centres de soins différents basés au Canada, aux États-Unis ou en Australie. En prenant leur point de vue, le vocabulaire suivant sera utilisé :

— **Client** ou **compagnie** : agence de soins à domicile qui utilise la plateforme d'AlayaCare
— **Employé** : employé d'un client
— **Patient** : personne qui reçoit un soin par un employé d'un client

L'objectif principal de cette étude est de développer un outil capable de créer un diagnostic

sur la satisfaction des employés d'une clinique de soins à domicile à partir des données d'activité professionnelle et d'améliorer un algorithme d'optimisation. Cet outil doit avoir un bon niveau d'explicabilité : l'importance de chaque variable utilisée doit pouvoir être relevée afin d'identifier les facteurs qui influencent la satisfaction des employés. De cette manière, on est capable de communiquer à une structure l'état de satisfaction globale de son personnel et les décisions qui peuvent avoir un effet positif sur les employés.

Le chapitre 2 propose une revue de littérature afin de mentionner les études qui ont déjà été menées en lien avec la mesure de la satisfaction des employés dans le domaine de la santé. Une approche méthodologique pour un projet d'exploration de données sera ensuite présentée au chapitre 3. Les différentes étapes nécessaires qui y sont détaillées seront ensuite appliquées à notre cas d'étude afin d'explorer la base de données d'AlayaCare dans le chapitre 4. Les choix de modélisation seront aussi présentés et discutés. Le chapitre 5 reprend les résultats des expériences, une analyse des performances des modèles et l'influence des variables explicatives. Enfin, le chapitre 6 sera consacré à l'applicabilité de nos résultats dans l'algorithme d'optimisation utilisé à AlayaCare, avec la proposition d'une nouvelle fonction de coût associé aux routes. Une conclusion reprenant une synthèse des travaux effectués, les limites de cette étude et des suggestions pour des travaux futurs clôturera ce projet.

CHAPITRE 2 REVUE DE LITTÉRATURE

L'objectif de cette revue de littérature est dans un premier temps d'étudier l'impact de l'insatisfaction des employés du domaine de la santé. De nombreux articles montrent des liens entre l'épuisement émotionnel (ou *burnout*), le problème de rétention d'employés, la qualité de soins et l'insatisfaction au travail. Les résultats d'études récentes cherchant à identifier les facteurs ayant un impact sur la satisfaction des employés sera ensuite présentée. Enfin, les méthodes d'autres articles effectuant une étude semblable seront présentées et discutées. L'objectif est ici de voir et comprendre les avancées et découvertes qui ont déjà été réalisées dans ce domaine afin de s'en servir comme base pour ce projet.

2.1 La mesure de la satisfaction des employés dans le domaine de la santé

De nombreuses études basent leur analyse sur des questionnaires et des sondages envoyés directement aux employés [5–13]. Généralement constitués d'une dizaine de questions, les réponses suivent souvent l'échelle de Likert, outil psychométrique permettant de mesurer le degré d'accord avec la question. Certaines études se sont penchées sur la question de la mesure de l'insatisfaction de l'emploi et de la volonté de quitter son travail [14–17], dont certains exemples seront présentés. Les résultats sont ensuite analysés avec des méthodes quantitatives : statistiques simples, mesures de corrélation ou en utilisant la régression logistique.

2.1.1 L'insatisfaction, la rétention des employés, et l'épuisement émotionnel

Ces trois notions ne sont pas indépendantes. Une étude réalisée par Judith Shiao en 2014 [6] apporte des conclusions sur l'importance de l'épuisement émotionnel dans la volonté de quitter son emploi. Une revue de littérature concernant la santé mentale chez les infirmières d'urgence [18] souligne la bi-directionnalité entre la rétention d'employés et l'épuisement émotionnel. Selon l'étude de Mikyoung Lee (2019) [19], la satisfaction de l'emploi est directement lié à la charge émotionnelle des employés à travers les interactions avec les patients. En s'intéressant aux préférences des infirmières en terme de longueur de quart de travail, Chiara Dall'Ora [5] affirme que les gestionnaires ont un impact direct sur la satisfaction des employés à travers la gestion des horaires, ce qui permet d'éviter l'épuisement émotionnel tout en élaborant de bonnes stratégies de rétention. Dans une optique d'évaluer les facteurs agissant sur la satisfaction au travail, les études concernant l'épuisement émotionnel et la rétention d'employés seront considérées comme pistes de solution. Ainsi, les facteurs ayant

un impact sur un de ces phénomènes pourra être réutilisé dans l'étude de la satisfaction des employés.

2.1.2 Outils de mesure de la satisfaction du personnel infirmier

À travers les études analysées dans ce chapitre, plusieurs outils de référence sont utilisés afin d'évaluer les intentions de départ des employés, la satisfaction au travail ou l'épuisement émotionnel. Trois instruments communément utilisés sont la *McCloskey/Mueller Satisfaction Scale* (MMSS), le *Nursing Working Index* (NWI) et le *Maslach Burnout Inventory* (MBI). S'intéresser à ces méthodes permet de voir les axes étudiés pour comprendre et quantifier la satisfaction des employés. En proposant une version plus condensée du MMSS, un article de 2006 [20] compare les résultats des trois tests afin de montrer les corrélations entre eux.

***Nursing Working Index* (1989)**

Développé à l'origine par Kramer et Hafner en 1989 [17], cette méthode a pour objectif de mesurer la satisfaction des infirmières dans différentes structures ou départements d'un point de vue de productivité et de qualité de soins. Les réponses sont constituées d'une échelle de quatre points pour évaluer les conditions de travail des infirmières. Plus tard révisé en *Revised Nursing Work Index* (NWI-R) et *Practice Environment Scale of the Nursing Work Index* (PES-NWI) par [21], le questionnaire est constitué de 49 questions séparé en 5 catégories : l'implication des infirmiers envers la structure, la qualité de soins, la qualité de l'encadrement par les directeurs, l'adéquation entre les employés et les ressources disponibles, et les relations entre les médecins et les infirmières. Certaines d'entre elles ne s'appliquent que dans le domaine hospitalier. Ces nouvelles versions ont diminué la pénibilité du test, tout en mettant l'accent sur l'environnement de travail pour apporter des bénéfices au personnel comme aux patients.

***McCloskey/Mueller Satisfaction Scale* (1990)**

La méthode de McCloskey et Mueller publiée initialement en 1974 et révisée dans sa version de 1990 [16] identifie trois dimensions pour mesurer la satisfaction des employés. La première concerne la sécurité, caractérisée par le salaire, les avantages liés au métier et la planification du travail. La deuxième dimension est centrée autour du social, en s'intéressant principalement aux relations avec les collègues de travail et les *managers*. La dernière dimension identifiée est l'aspect psychologique de la satisfaction au travail, composée entre autres de la reconnaissance reçue au travail, des responsabilités liées au métier et des opportuni-

tés. Cet outil a été largement utilisé dans des contextes géographiques et professionnels très différents [20].

Le *Maslach Burnout Inventory* (1997)

Un outil mondialement reconnu et utilisé pour mesurer l'épuisement au travail est le MBI [14]. L'épuisement émotionnel et la satisfaction au travail sont deux phénomènes reliés [6, 19]. Les facteurs explicatifs du *burnout* sont divisés en trois catégories : l'épuisement émotionnel, la dépersonnalisation, et la réalisation personnelle. L'autrice souligne que dans le domaine des soins, un employé est sans cesse confronté à des patients en quête de solutions à leurs problèmes (psychologiques, sociaux ou physiques). Ceci peut s'accompagner de sentiments négatifs comme du stress, de l'embarras, de la peur. Cette charge émotionnelle peut parfois être trop lourde à supporter, et mener à une surcharge émotionnelle. Les principales conséquences d'une dégradation de la santé mentale ne concerne pas seulement l'employé, mais aussi la qualité de soins qui sera donné au patient [10, 14, 19, 22]. Ce questionnaire est constitué de 22 questions axées sur la mesure d'une des catégories susnommées. Un score pour chaque catégorie est estimé. La méthode présente le cadre dans lequel le questionnaire doit être administré afin que les réponses puissent être considérées comme valides.

2.1.3 Impact sur la qualité de soins

Certaines études se sont focalisées sur les impacts de l'épuisement au travail, l'insatisfaction et la frustration sur les conditions de travail. L'article de McHugh publié en 2011 [10] met en avant l'écart de satisfaction mesuré entre le personnel qui pratique les soins de manière directe (cliniques, hôpitaux) et le personnel du milieu pharmaceutique. Dans cet article, il est montré qu'un haut niveau d'épuisement chez les infirmières est lié à une forte insatisfaction des patients. Ainsi, l'étude de la satisfaction des patients peut apporter de l'information sur la satisfaction des infirmières.

Une revue de littérature menée par Sung-Heui Bae [22] se concentre sur l'impact des heures de travail des infirmières, notamment les dépassements horaires sur les conditions des patients. Certains articles relèvent qu'une mauvaise gestion du dépassement horaire peut conduire à une augmentation des erreurs médicales ou à une baisse de satisfaction des clients. La gestion des horaires par un gestionnaire a donc indirectement un impact sur la qualité de soins donnés aux patients et donc sur la satisfaction des employés. Améliorer le niveau de satisfaction des employés pourrait donc avoir un impact positif sur la qualité de soins offert aux patients. Cette intuition est soutenue par [19] qui souligne l'impact que peuvent avoir les *managers* sur le bien-être de ses employés et les retombées directes pour la structure de soins en question

et la qualité de soins offerte aux patients.

2.2 Facteurs explicatifs de la satisfaction des employés

À travers la lecture de la littérature, on est capable de distinguer 5 familles de facteurs ayant un impact sur la satisfaction des employés : les caractères propres à l'individu, ceux en lien avec la planification du travail, les caractéristiques du métier, les interactions sociales, et le rôle de la structure dans lequel l'employé travaille.

Le statut du personnel de santé des études présentées dans ce chapitre est susceptible de varier. Certains articles décident de se concentrer sur un type en particulier : infirmiers hospitaliers, infirmiers en soins à domicile, ou infirmiers dans des équipes d'urgence. Dans la base de données fournie par AlayaCare, notre partenaire industriel fournisseur de données, les agences possèdent du personnel qui peut être associé ponctuellement à des créneaux qui s'apparentent à du travail hospitalier. Ainsi, il apparaît important d'explorer les variables de tous ces milieux qui partagent beaucoup de points communs afin d'apprécier la généralisation des algorithmes présentés dans cette étude et de mesurer l'impact de certaines variables dans les différentes catégories.

2.2.1 Indicateurs individuels et démographiques

On trouve dans la littérature de nombreux facteurs de la vie personnelle d'un individu ayant des impacts sur la satisfaction au travail. Une étude de 2020 [11] en souligne plusieurs. Les personnes de plus de 36 ans semblent moins affectées par le *burnout syndrome*. Les niveaux de satisfaction des femmes sont généralement plus élevés. La même observation est faite pour les personnes blanches parlant couramment anglais. De plus, la consommation d'alcool et de cigarettes peuvent avoir des effets négatifs tandis qu'une activité physique régulière semble améliorer le confort de vie des employés. La qualité de sommeil peut jouer un rôle dans l'apparition du *burnout syndrome* [12]. La durée de sommeil ainsi que sa qualité peuvent être amenés à décroître dans ce cas, ce qui peut amener à des abandons professionnels.

Karsh [23] aboutit à des conclusions semblables : les personnes les plus âgées, de sexe féminin, blanches et parlant couramment anglais semblent montrer de meilleurs niveaux de satisfaction. Cependant, une étude spécifique sur des employés en soins à domicile [8] montre que l'âge n'a que très peu d'influence sur l'intention de quitter son emploi. Les conclusions peuvent varier d'un contexte à un autre.

2.2.2 Planification du travail

Certains facteurs sont reliés à la planification du travail du personnel soignant, en particulier les horaires de travail : le dépassement horaire a des incidences négatives sur les employés [6, 22]. Associé à plus de fatigue et moins de concentration, il contribue à l'apparition du *burnout syndrome*. L'objectif généralement cité se situe autour de 40 heures par semaine pour une personne employée à temps plein, mais cette valeur peut varier selon les contextes ou les pays. De manière analogue, travailler trop peu est associé à des niveaux de stress plus élevés et peut mener au même syndrome [18]. Il est aussi recommandé de limiter l'exposition à des évènements traumatiques comme des patients dont l'état s'est fortement dégradé ou des patients suicidaires par exemple.

Un autre aspect important est l'arrangement des horaires de travail. Le travail de nuit peut être un facteur de risque [11], et une bonne gestion des temps de pause est essentielle. Les distances parcourues, en particulier en soins à domicile est un facteur de pénibilité [8]. De manière générale, prendre en compte les préférences des employés dans la création des emplois du temps est très important pour maximiser leur satisfaction.

2.2.3 Caractéristiques liées au métier

Les postes occupés par les employés de soins et les différents avantages qui les accompagnent ont une influence sur la satisfaction. On retrouve différents types de personnels infirmiers, plus ou moins spécialisés et demandant une formation plus ou moins longue. Les niveaux de satisfaction peuvent varier selon la catégorie d'emploi occupée [7]. Les postes qui impliquent des soins directs aux patients sont associés à des risques de *burnout syndrome* plus élevés. Certains avantages comme la couverture sociale ou le plan de retraite sont aussi des facteurs importants au sein des employés de soin. Ils s'accompagnent de plus de sécurité et donc contribuent à la satisfaction des employés [10]. Il faut aussi distinguer les postes managériaux du personnel régulier, qui sont généralement mieux rémunérés et ont plus d'avantages [23], car un bon salaire est une source directe de satisfaction [6, 8].

2.2.4 Interactions sociales

Le personnel de soins interagit principalement avec trois types de personnes : les patients, les collègues de travail, et les superviseurs/*managers*. Voir des patients qui nécessitent des soins à longueur de journée entraîne une charge émotionnelle forte [14]. Un comportement superficiel (*surface acting*) avec les patients est corrélé avec des émotions négatives, tandis qu'un comportement plus naturel (*deep acting*) est lié avec du plaisir et de la fierté [19]. Avoir

de bonnes relations avec ses collègues de travail semble aussi avoir un impact positif sur la satisfaction au travail [24]. Les rapports entre les superviseurs et les employés permettent de créer un climat de confiance nécessaire à la satisfaction des employés [13, 15, 18]. Le rapport direct et humain est à privilégier, les interfaces électroniques créant de la distance [4].

2.2.5 Rôle de la structure

Certaines caractéristiques dépendent de la structure du personnel soignant. Celle-ci doit garantir une certaine autonomie aux employés, ainsi qu'un bon environnement de travail afin de prévenir l'apparition du *burnout syndrome* [18]. La formation du personnel est un élément qui ne doit pas être négligé [4]. Un employé avec une bonne connaissance des tâches qu'il doit réaliser et des clients qu'il doit rencontrer s'accompagne de meilleurs résultats et un plus grand attachement à la structure de soins. Lors de ces visites, les ressources techniques telles que l'équipement ou le matériel doivent être suffisantes et de bonne qualité [11]. L'emplacement géographique de l'agence semble aussi impacter la satisfaction globale des employés. Il a été relevé que les niveaux de satisfaction étaient plus élevés dans les petites zones urbaines [7].

2.3 Modèles quantitatifs utilisés

L'exploration d'articles réalisée dans le cadre de cette étude n'a pas permis de trouver de méthodes quantitatives se basant directement sur des données autres que des réponses à des questionnaires ou des sondages. Avoir des modèles qui peuvent exploiter la notion temporelle des données fournies par AlayaCare, notamment les données des visites des employés semble nécessaire afin d'identifier les facteurs ayant un impact sur la satisfaction des employés et leur intention de quitter leur emploi. On est tout de même capable d'identifier certaines méthodologies appliquées sur des employés de soins. D'autres études se sont penchées sur la rétention des patients dans le domaine de la santé, et non des employés. Dans les deux cas, il s'agit d'évaluer la stabilité de la condition d'une personne affiliée à une compagnie ou une structure. Au vu de la similarité entre les deux problèmes, on pourra faire l'hypothèse que les modèles appliqués dans une situation peuvent être adaptés à l'autre.

2.3.1 Régression logistique

La régression logistique est un modèle linéaire généralisé qui permet de déterminer les coefficients $b_0, ...b_n$ des variables explicatives $x_1, ..., x_n$ permettant de décrire au mieux une variable

y selon le modèle suivant [25] :

$$P(y = 1|X) = \frac{e^{b_0+b_1x_1+...+b_nx_n}}{1 + e^{b_0+b_1x_1+...+b_nx_n}}$$

Dans une régression logistique simple, la variable à expliquer y est binaire. Pour [23] par exemple, l'objectif est de classifier les employés qui ont manifesté leur désir de partir par rapport à ceux qui préfèrent rester dans leur structure de soin actuelle. [26] effectue la même classification. [27] présente une manière d'ajouter de la régularisation aux modèles, en utilisant les normes L_1 et L_2. Cette méthode permet d'annuler les coefficients des variables qui n'apportent pas ou très peu d'information dans la prédiction en ajoutant un terme à la fonction de vraisemblance. La librairie `scikitlearn` propose une version hybride qui utilise les deux normes avec un paramètre `l1_ratio` permettant de définir l'importance de l'une sur l'autre. Une telle régularisation est appelée *elasticnet*.

2.3.2 Arbres de classification et forêts aléatoires

Les arbres de classification sont des méthodes non-paramétriques qui permettent de sectionner l'espace des variables explicatives en zones rectangulaires [28]. L'objectif est de grouper dans chaque zone des observations appartenant à une même classe. Le nombre de découpages et la profondeur de l'arbre doivent être sélectionnés afin d'éviter le phénomène de sur-apprentissage. L'inconvénient majeur de cette méthode est l'instabilité des frontières de décision : une petite modification des données peut changer complètement la prédiction de classe d'une instance.

La méthode des forêts aléatoires [28] permet d'assouplir la sensibilité des données. Cette méthode se décompose de la manière suivante :

- Créer des ensembles d'apprentissages différents en échantillonnant les observations et les variables explicatives aléatoirement
- Entraîner un arbre de décision sur chaque échantillon
- Déterminer la classe d'un exemple en regardant la prédiction majoritaire parmi l'ensemble des arbres de décision

Cette méthode a été utilisée par [26] dans le contexte de rétention de clients en se basant sur des données temporelles. L'objectif de classification est le même que pour la régression logistique : les individus considérés comme stables, et ceux qui ont quitté leur service. L'inconvénient principal lié à cette méthode est qu'il n'y a pas d'arbre de décision unique, il n'est donc pas évident de visualiser le modèle final. Cependant, il existe plusieurs métriques afin de relever l'importance des variables. Une méthode naïve est de compter le nombre de fois qu'une

variable est utilisée par un noeud de l'arbre. La méthode implémentée par `scikitlearn` est présentée par [28] : *permutation accuracy importance*. Cette méthode évalue l'importance de chaque feature en mélangeant les valeurs pour observer l'effet entraîné sur la prédiction. On obtient alors l'importance relative de chaque variable explicative, la somme de tous les scores d'importance étant égale à 1.

2.3.3 Méthodes de *boosting* et XGBoost

Les méthodes de *boosting* sont de plus en plus utilisées en apprentissage machine [29]. L'objectif est de construire un classificateur fort à partir de classificateurs faibles. À chaque itération, un modèle de base est crée afin de corriger les erreurs du modèle principal, jusqu'à ce que les performances n'évoluent plus. Ce processus est séquentiel : les itérations sont successives. Le *gradient boosting* se caractérise par l'utilisation de l'algorithme de la descente du gradient afin de minimiser la fonction de perte défini dans le problème de classification ou de régression.

L'algorithme XGBoost disponible dans différents langages informatiques est basé sur le principe de *gradient boosting*. Il est appliqué sur des arbres de décisions et est optimisé en vitesse et en performance. Ce modèle est largement utilisé dans des problèmes de classification populaires, il atteint notamment des premières places dans de nombreuses compétitions proposées par le site *Kaggle*. Il est entre autres utilisé dans un problème de rétention de clients avec des données temporelles [30].

2.3.4 Réseaux de neurones

Les réseaux de neurones constituent un autre type d'architecture largement utilisé. La structure *Multi Layer Perceptron* (MLP) possède une couche d'entrée, des couches intermédiaires ainsi qu'une couche de sortie [31]. Les variables explicatives sont entrées dans le modèle à travers la couche d'entrée. Chaque neurone des couches cachées reçoit une somme pondérée par le vecteur de poids w associé aux sorties des neurones de la couche précédente x, puis applique une fonction d'activation g. La valeur produite z est ensuite utilisée par la couche suivante :

$$z = g(w^T x + \omega_0)$$

Les poids initiaux sont choisis aléatoirement, puis leur valeur est mise à jour à chaque itération afin de minimiser une fonction de perte. Les fonctions d'activation g couramment utilisées sont la fonction tangente hyperbolique et la fonction sigmoïde. Augmenter le nombre de couches cachées permet d'aller chercher des relations plus profondes, mais suivre l'information afin d'identifier quelle variable explicative a eu de l'impact sur la variable de sortie est rendu

beaucoup plus dur. Ainsi, par rapport aux autres méthodes, l'explicabilité de la variable de sortie et la sélection de variables explicatives est moins adapté. Ce type d'architecture a déjà été utilisé dans un problème de classification binaire : la réadmission de patients dans un hôpital [32].

2.3.5 Méthodes de balancement des classes

Une étude du comportement des modèles de classification avec des méthodes de sur-échantillonnage et de sous-échantillonnage dans des problèmes classiques avec des classes non balancées a été réalisée en 2014 [33]. Deux méthodes de sur-échantillonnage sont utilisées : le sur-échantillonnage aléatoire et la méthode SMOTE [34]. Cette dernière méthode consiste à générer des observations synthétiques en utilisant le voisinage des instances de la classe minoritaire. Parmi les méthodes de sous-échantillonnage, l'article présente le sous-échantillonnage aléatoire, les liens Tomek ainsi que la méthode *Edited Nearest Neighbours* (ENN). La méthode des liens Tomek propose d'enlever des paires de points lorsque ceux-ci sont mutuellement plus proche voisin et appartiennent à une classe différente. On peut alors faire l'hypothèse qu'ils sont proches d'une frontière de décision. La méthode ENN prend en compte le voisinage des instances d'un jeu de données afin de valider qu'une majorité des voisins appartiennent à la même classe que le point concerné. Les différentes combinaisons de ces méthodes sont testées sur plusieurs jeux de données classiques, puis comparés. Selon les cas, certaines méthodes performent mieux que d'autres, il est donc pertinent d'explorer ces méthodes afin de trouver celle qui correspond le plus à un cas d'étude spécifique. Combiner certaines stratégies de sur-échantillonnage et de sous-échantillonnage sont proposées par [33] tel que :

- Sur-échantillonnage SMOTE puis sous-échantillonnage Liens Tomek
- Sur-échantillonnage SMOTE puis sous-échantillonnage ENN

Les facteurs sources de satisfaction pour les employés de soins à domicile sont multiples. Certains d'entre eux sont accessibles au cours du temps à travers la base de données de notre partenaire industriel. À partir des outils mathématiques présentés dans cette section, on espère pouvoir vérifier les hypothèses émises dans la littérature afin de proposer un modèle capable d'identifier les causes de l'insatisfaction au travail et de l'intention de quitter son emploi.

CHAPITRE 3 MÉTHODOLOGIE

Dans tout projet d'exploration de données, il est nécessaire d'avoir une méthodologie précise et rigoureuse. Dans ce projet, il a été choisi de suivre la méthode *Cross Industry Standard Process for Data Mining* (CRISP-DM) [35]. Il s'agit d'un modèle hiérarchique qui définit pas-à-pas les étapes qu'il faut respecter dans un projet d'exploration de données. Les quatre étapes principales sont les suivantes :

— la compréhension des données
— la préparation des données
— la modélisation
— l'évaluation

Des explications de chacune de ces étapes sont présentées dans les sections suivantes. Il faut noter que la méthode présentée dans l'ensemble de ce chapitre est largement inspirée de [35].

3.1 La compréhension des données

3.1.1 Collecte de données

La première étape dans un projet d'exploration de données est d'acquérir les données. Il faut être en mesure d'extraire les informations intéressantes pour l'étude, à travers l'ensemble des données disponibles à la source. Par exemple, on doit déterminer quelles périodes dans l'historique de données sont nécessaires, même si cela implique de ne pas exploiter la totalité de la base de données.

Les données peuvent être agrégées dans différentes tables, il sera alors nécessaire de faire des jointures. Ce procédé peut être une source d'erreur si les clés qui permettent de faire les jointures ne sont pas présentes dans les deux tables ou ne sont pas uniques par exemple. On a alors une perte de données, ou des données inutilisables. Les variables doivent être sélectionnées sous le prisme des hypothèses qui ont été formulées pour le problème d'exploration de données.

3.1.2 Description des données

L'étape de compréhension des données est essentielle avant de manipuler l'information. La source des données et la manière dont les informations ont été collectées doivent être maîtrisées. Certaines données ont pu être entrées manuellement, ou automatiquement comme les données d'un capteur électronique par exemple. Dans les deux cas, l'analyse d'erreur ne sera

pas la même. Par ailleurs, la variable cible de l'étude doit être présente afin de pouvoir valider les hypothèses formulées en début de projet. Une analyse préliminaire permet d'apprécier le type et l'amplitude de chacune des variables. Celles-ci pourront par exemple être :

- numériques : valeur continue ou discrète
- binaires : il s'agit d'une valeur qui peut être vraie ou fausse uniquement, comme avoir le diabète par exemple
- catégoriques : un nombre fini de catégories, comme le sexe d'une personne par exemple
- libre : champ texte libre, les attributs peuvent prendre n'importe quelle valeur

Une fois l'origine des données correctement identifiée et comprise, l'étape suivante consiste à explorer les valeurs des différentes variables.

Analyse d'une variable

Chaque variable peut être explorée individuellement. La distribution permet de voir la répartition d'une valeur au sein de la population. Certaines caractéristiques statistiques de base peuvent être relevées comme la moyenne, la variance, les valeurs minimales et maximales, et le mode (valeur la plus fréquente). Ceci permet entre autres de faire des premières observations sur une variable :

- la dispersion des valeurs pour l'ensemble de la population : si une variable est constante, elle ne permet pas de distinguer les individus
- la répartition des valeurs : on peut regarder notamment comment se comportent les valeurs extrêmes par rapport au reste de la distribution
- des schémas caractéristiques, ou des irrégularités dans les données qui pouvaient être anticipées ou non. Par exemple, dans un jeu de données sur du personnel infirmier, on s'attend à observer plus de femmes que d'hommes.

Analyse de plusieurs variables entre elles

Lorsqu'on observe un jeu de données sous plusieurs dimensions, il est nécessaire d'observer les relations entre ces variables. On peut par exemple regarder les corrélations entre deux variables. Le coefficient de corrélation linéaire de Pearson entre une variable explicative X et la variable cible Y est défini comme suit :

$$r_{X,Y} = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}{\sigma_X \sigma_Y}$$

avec σ l'écart type d'une variable aléatoire. Ce facteur permet de voir s'il y a une corrélation linéaire entre les deux variables. Si la valeur de ce coefficient est proche de zéro, cela veut

dire que les variables ne sont pas corrélées linéairement, mais peuvent cependant l'être de manière non-linéaire (quadratique, exponentielle, ou autre). On retrouve dans le tableau 3.1 les valeurs du coefficient qui permettent de classifier si la corrélation est forte ou faible, positive ou négative.

Tableau 3.1 Classification du coefficient de corrélation

Corrélation	Positive	Négative
Faible	]0 ; 0,5]	[-0.5 ; -0[
Forte	]0.5 ; 1]	[-1 ; -0.5[

3.1.3 Validation des données

Les données ayant été sélectionnées pour leur intérêt dans le projet, il faut passer par une étape de validation. La présence de biais dans les données risque de fausser l'interprétation des résultats ou de mener vers des mauvaises conclusions sur les hypothèses de départ. La cohérence des données entre elles doivent avoir du sens : par exemple, il faut être capable de détecter si un évènement se termine avant d'avoir commencé. Il faut comprendre la source d'éventuelles incohérences afin de pouvoir corriger l'information. Si la compréhension n'est pas suffisante dans un cas, il est parfois préférable ne pas considérer l'information.

Lorsqu'on travaille avec une variable catégorique, il faut valider que chaque catégorie est distincte et est intéressante pour l'étude réalisée. Si la source de données est bilingue et contient une variable pour le sexe d'un individu, il faut regrouper les catégories *Homme* et *Male*, ou *Other* et *Unknown* par exemple.

Travailler avec des variables avec champ de texte libre implique de vérifier le nombre d'entrées distinctes. La présence de majuscules, d'abréviations ou d'erreurs de saisies créent des valeurs différentes alors qu'elles apportent la même information. Utiliser des expressions régulières peut être un moyen de classifier un ensemble de valeurs. Par exemple, l'expression régulière « $\hat{}$*infirm* »permet de regrouper les mots «infirmière », «infirmier », «infirmière auxiliaire», ou toute chaîne de caractère commençant par «infirm ». Supprimer les accents et les majuscules est une bonne étape préliminaire pour faire une recherche robuste. Une fois les catégories principales analysées, le reste des valeurs non classifiées peuvent être ajoutées à une catégorie par défaut, ou considérées comme manquantes.

3.2 Préparation des données

3.2.1 Restriction des données

L'étape d'exploration aura éventuellement permis de cerner les variables d'intérêt. L'étape de validation précédente permet de construire une liste d'hypothèses et d'actions à appliquer sur les données afin qu'elles soient utilisables. Une nouvelle étape de sélection peut donc être réalisée sur certains attributs afin de restreindre l'étude à la population d'intérêt. Dans cette étape, il est possible que certaines informations soient abandonnées.

Selon les modèles utilisés, il est important de regarder le balancement des classes, particulièrement dans un problème de classification. Des modèles comme les régressions logistiques sont sensibles à la proportion des classes dans l'estimation des coefficients. Deux techniques principales sont souvent envisagées :

— Sous-échantillonnage : l'objectif est de réduire le nombre d'exemples de la classe majoritaire. Le sous-échantillonnage peut être réalisé de manière aléatoire, ou de manière méthodique en utilisant les liens Tomek [36] par exemple.

— Sur-échantillonnage : l'objectif est d'augmenter le nombre d'exemples de la classe minoritaire. Il est possible d'utiliser des techniques aléatoires qui dupliquent certaines données ou de générer des observations synthétiques en utilisant la méthode SMOTE [34] par exemple.

Des combinaisons de ces deux types de méthodes peuvent donner de bons résultats, il faut alors explorer comment les utiliser au mieux pour avoir les meilleurs résultats possibles. Pour les méthodes de sur-échantillonnage, il est prudent de procéder au nettoyage des données présenté dans la section suivante avant de générer des observations synthétiques afin d'éviter d'amplifier des erreurs.

3.2.2 Nettoyage des données

À cette étape, les correctifs qui ont été précédemment identifiés après l'analyse des différentes variables peuvent être mis en place. Il faut alors choisir de corriger certaines données, et décider d'en enlever d'autres. Un problème récurrent est la proportion de données manquantes. S'il y a des valeurs manquantes pour une variable, les causes doivent être identifiables grâce au travail de compréhension réalisé en amont. De cette manière, on est en mesure de réaliser un traitement adéquat pour cette variable. Plusieurs stratégies peuvent être menées afin de pallier au problème des données manquantes :

— Mettre les valeurs manquantes à une valeur par défaut (0 par exemple si la variable est numérique)

— Mettre les valeurs manquantes à la valeur moyenne ou la valeur la plus fréquente de cette variable
— Modéliser la variable avec une loi connue puis réaliser des tirages aléatoires
— Retirer les exemples lorsque la donnée est manquante
— Retirer la variable du jeu de données si la proportion de données manquantes ou erronées est trop importante

3.2.3 Construction des données

Transformation d'une variable

Une pratique courante est de normaliser les données. Cette pratique permet aux modèles de ne pas tenir compte des différences d'amplitude entre les variables (comme l'âge et le revenu par exemple [35]). Une transformation largement utilisée est la transformation en une loi de moyenne nulle et d'écart type unitaire. Pour ce faire, il faut appliquer la formule suivante à la variable aléatoire X :

$$X' = \frac{X - \mu}{\sigma}$$

avec μ l'espérance de X et σ son écart type. Ainsi, X' vérifie $\mathrm{E}(X') = 0$ et $\mathrm{Var}(X') = 1$. Il faut noter que certains modèles sont insensibles à une telle transformation (modèles d'arbres par exemple). Si de nouvelles observations sont utilisées pour de l'entraînement ou des prédictions, le même pré-traitement doit être réalisé afin que le modèles prenne des entrées cohérentes. Cette transformation n'est pas nécessaire si la variable est binaire ou catégorique.

Transformation de plusieurs variables

Deux variables peuvent être combinées pour créer une nouvelle information. Celle ci peut être par exemple le produit ou le rapport de deux autres variables, dont l'information n'aurait pas pu être captée par des modèles linéaires. Il faut cependant rester prudent, car il y a un risque de créer des corrélations fortes entre les variables explicatives. Le rang d'une matrice de variables explicatives peut alors diminuer et ne plus être inversible, ce qui peut avoir un impact négatif sur la convergence de certains algorithmes, comme les régressions linéaires par exemple.

3.3 Modélisation

Le type de problème ayant été identifié (prédiction, classification, génération par exemple), des modèles appropriés peuvent être sélectionnés, comme ceux présentés à la section 2.3. Il

est fortement conseillé d'essayer différentes architectures afin de valider les modèles produits. Des architectures différentes peuvent saisir des relations différentes : linéaires, ou non linéaires par exemple. Les résultats des différents modèles peuvent ensuite être croisés pour valider les conclusions des expériences.

Dans ce type de projet, il est courant de séparer l'ensemble des données en trois jeux distincts : l'ensemble d'entraînement, de validation, et de test [37]. Le premier, comme son nom l'indique, est utilisé pendant l'entraînement. Pour une régression logistique par exemple, les coefficients attribués à chaque variable seront déterminés à partir de cet ensemble. Le second jeu de données va être utilisé pour la validation : pour un type de modèle, de nombreux hyper-paramètres doivent être essayés. Les paramètres optimaux peuvent être sélectionnés en observant les performances sur l'ensemble de validation [37]. Des métriques adaptées au problème doivent être déterminées afin de valider les modèles choisis à cette étape. L'ensemble de validation ayant été utilisé plusieurs fois, le troisième jeu de données (l'ensemble de test) ne sera appliqué que sur le modèle final afin d'observer les performances sur un ensemble d'exemples qui n'a pas été observé pendant le processus de sélection de modèle [37]. Cette méthodologie est cruciale afin de pouvoir évaluer si un modèle fait du sur-apprentissage.

Dans une tâche de classification binaire, on peut s'intéresser à la matrice de confusion du tableau 3.2. Les termes de vrai/faux positifs/négatifs y sont présentés. La variable la plus couramment utilisée dans les modèles de classification est le taux de bonnes prédictions :

$$\text{Taux de bonnes prédictions} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}}$$

On peut aussi définir la précision, le rappel et la F1-Score F_1 de la classe positive de la manière suivante :

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

$$F_1 = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Contrairement au taux de bonnes prédictions, l'avantage principal des trois dernières métriques est qu'elles permettent de prendre en compte le caractère non équilibré des classes.

La métrique F_1 doit être comparée avec les performances d'un modèle de classification naïf. Soit c la proportion de la classe positive tel que $P(y = 1) = c$. En prenant un modèle qui

prédit toujours la classe positive, alors le rappel vaut 1 et la précision vaut c. On a alors

$$F_{1,naïf} = 2 \cdot \frac{c}{c+1}$$

La F_1 d'un modèle qui performe bien sera proche de 1 et le plus éloigné possible du $F_{1,naïf}$. Une manière de visualiser cette information est de regarder la métrique suivante :

$$\text{Signal}_{\%} = 100 \cdot \frac{F_1 - F_{1,naïf}}{1 - F_{1,naïf}}$$

Cette métrique est une échelle linéaire entre les performances d'un algorithme naïf et celles d'un algorithme parfait. En effet celle-ci vaut 0 lorsque la F_1 obtient les mêmes performances qu'un algorithme naïf. Elle vaudra 100 si la performance maximale est obtenue. Si les performances obtenues sont moins bonnes que celles d'un algorithme naïf, elle sera alors négative. Ainsi, elle peut être interprétée comme le pourcentage de signal perçu par une solution.

Tableau 3.2 Matrice de confusion

		Réalité	
		Positif	Négatif
Prédiction	Positif	Vrais positifs (VP)	Faux positifs (FP)
	Négatif	Faux négatifs (FN)	Vrais négatifs (VN)

Une autre métrique fréquemment utilisée dans les problèmes de classification est la courbe de la fonction d'efficacité du récepteur, plus communément appelée courbe ROC. La courbe peut être interprétée comme l'évolution du taux de vrais positifs lorsqu'on fait évoluer la valeur du seuil de classification (seuil entre 0 et 1 au delà duquel la valeur prédite est considéré comme un exemple positif). La valeur de l'aire présente sous cette courbe ROCAUC (*Receiver Operating Characteristic, Area Under the Curve*), varie entre 0 et 1 et permet d'évaluer la force du classificateur. Une valeur de 1 montre un classificateur parfait, tandis qu'une valeur de 0,5 correspond à un classificateur aléatoire.

3.4 Évaluation et discussion autour des résultats

Une fois les modèles sélectionnés, on peut procéder à l'analyse des résultats. Les interprétations peuvent être différentes d'une architecture à l'autre, comprendre les effets de chacune est essentiel. Les résultats peuvent ensuite être utilisés pour ordonner les meilleurs modèles. Les hypothèses formulées en début de projet peuvent être validées ou rediscutées. Une discussion argumentée autour des résultats permet de vérifier que les objectifs ont été atteints

et éventuellement proposer de nouvelles pistes de recherche suite aux conclusions proposées.

CHAPITRE 4 APPLICATION DE LA MÉTHODE

Dans ce chapitre, la méthode présentée au chapitre 3 est appliquée sur le jeu de données fourni par AlayaCare. Les étapes de compréhension des données, de préparation des données et de modélisation y sont détaillées.

4.1 La compréhension des données

Fondée en 2014, l'entreprise AlayaCare est le partenaire industriel qui nous fournit ses données. L'entreprise propose une plateforme de gestion pour des centres de soins domicile. À travers leur application, les gestionnaires ont accès à la liste des employés et des patients et créer des emplois du temps en créant des visites. Chaque employé reçoit sur son cellulaire la route qu'il doit suivre, et celui-ci peut rentrer des informations supplémentaires sur le déroulement d'une visite. Il faut prendre en compte plusieurs facteurs :

- les clients d'AlayaCare sont principalement répartis au Canada, aux États-Unis et en Australie. Les systèmes de soins étant différents, les facteurs influençant la satisfaction au travail peuvent varier selon le cas étudié
- le nombre d'employés dans une compagnie, s'il est élevé, assure un bon nombre d'exemples pour les jeux de données d'entraînement
- le nombre de visites au sein d'une compagnie qui dépend de l'ancienneté et de la taille de la compagnie permet de développer des variables explicatives riches
- les employés des clients peuvent avoir des formations et des emplois variés, même au sein d'une même structure

Les données de 117 clients ont été transmises, ce qui représente 12 millions de visites pour 63533 employés enregistrés dans le système d'AlayaCare.

4.1.1 Origine des données

Pays d'origine des clients

Avec deux implantations au Canada (Montréal et Toronto) et une en Australie (Sydney), la provenance des clients dont les données ont été extraites est détaillée dans le tableau 4.1.

Tableau 4.1 Origine des clients d'AlayaCare

Pays d'origine	Nombre de clients
Canada	72
États-Unis	24
Australie	21
Total	117

Système d'AlayaCare

L'entreprise AlayaCare propose une plateforme d'aide à la planification de visites et à la gestion administrative pour les agences de soins à domicile. L'utilisateur principal est le coordonnateur ou le gestionnaire qui affecte les patients à des employés via l'interface du logiciel. Les employés peuvent avoir un suivi de leurs visites sur une application mobile, et peuvent rentrer des informations sur l'état de la visite, comme les heures de début et de fin ou des commentaires divers.

4.1.2 Description de la structure de données

Les données sont stockées en format *.csv*, et chaque fichier représente une table de données pour un client. Les différentes tables disponibles, leurs attributs et les différents liens entre elles sont présentés dans la figure 4.1. Lorsque de l'information doit être partagée d'une table à une autre, on effectue une jointure sur les clés correspondantes.

La majorité des informations qui nous intéressent se retrouvent dans deux tables : la table Visites et Employés. Les données ayant été anonymisées pour le respect de la vie privée des clients d'AlayaCare, certaines informations sont disponibles sous la forme d'un identifiant numérique. Le nom des employés et des patients, les types de soins, les installations dans lesquelles ont eu les visites font partie des colonnes anonymisées. Les adresses sont données sous la forme d'un couple latitude/longitude avec une précision suffisamment faible pour ne pas pouvoir reconnaître un lieu spécifique ou identifier un lieu de résidence. Les clients dont les données ont été extraites ont accepté de partager leurs données dans un objectif de recherche.

Certaines valeurs sont entrées en champ libre de texte. Ceci concerne notamment les commentaires, le genre d'un employé, et le titre de son métier. D'autres comme le statut des employés, le type de disponibilité ou le statut d'approbation et d'annulation d'une visite sont catégoriques. Tous les identifiants sont numériques.

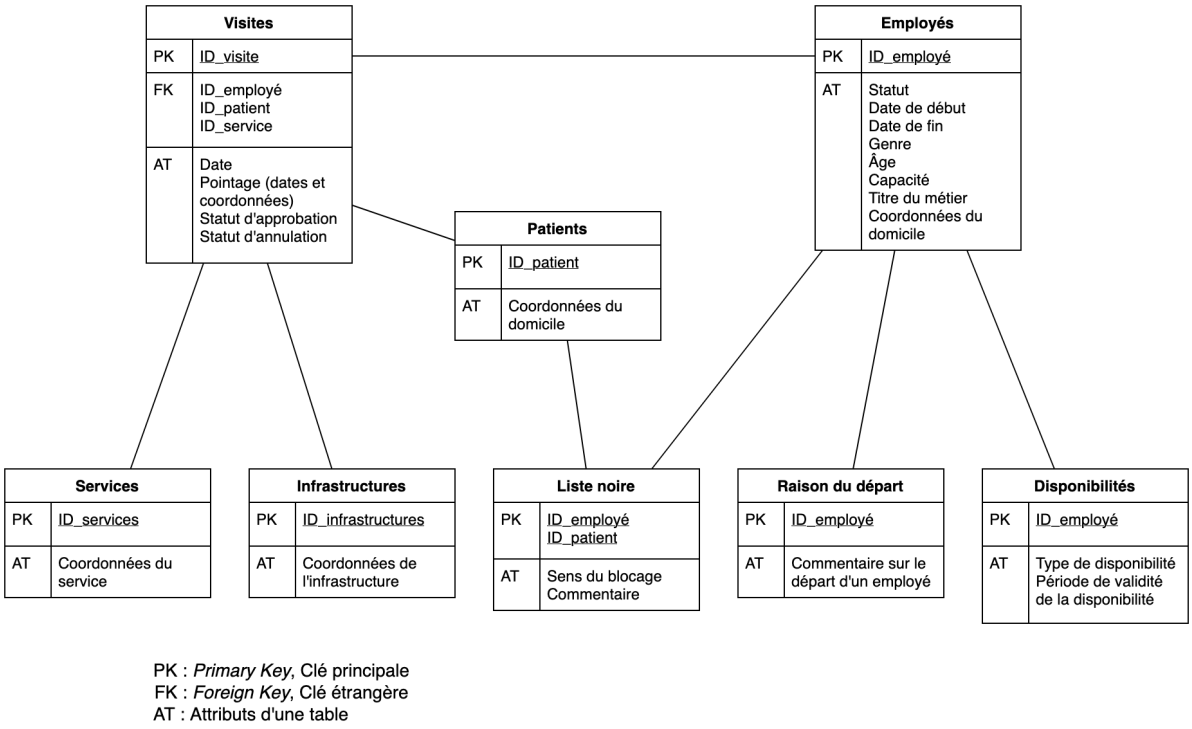

Figure 4.1 Structure des données

4.1.3 Modifications de la table Visites

La table principale concerne les données des visites. On y retrouve les informations suivantes :

- l'identifiant de l'employé associé à la visite
- l'identifiant du patient associé à la visite
- le type de soins donné
- des informations si la visite a eu lieu dans une installation particulière
- la date et l'heure du début de la visite, ainsi que de la fin estimée
- le statut d'approbation
- le statut d'annulation, avec éventuellement un code indiquant la raison
- les informations sur le pointage de la visite par l'employé au début et à la fin de la visite (heure et coordonnées géographiques)

Cette table est mise à jour en continu au fil de l'utilisation par un client. Si l'information d'une visite change, comme l'identifiant du personnel assigné, le statut d'approbation, ou la date et heure de la visite par exemple, le gestionnaire peut venir modifier manuellement le champ concerné.

Filtrage sur le statut d'un employé

L'information du statut des employés assignés aux visites de la table VISITES est obtenu en effectuant une jointure avec le statut des employés de la table EMPLOYÉS. Les quatre valeurs de statuts possibles pour un employé sont :

- *Active*, lorsque l'employé actif
- *Terminated*, lorsque l'employé ne travaille plus avec ce client
- *On-Hold*, lorsque l'employé est en pause, pour un congé maternité ou de longues vacances par exemple
- *Admin*, lorsque l'employé ne fait pas partie du personnel de soins

Dans le cadre de cette étude, seuls les employés *Active* et *Terminated* sont conservés. Le label *On-Hold* est ambigu : il est possible qu'un employé ayant quitté son travail revienne plus tard, le gestionnaire a donc le choix de le maintenir dans la base de données avec ce statut. Ainsi, pour éviter d'introduire des incertitudes dans les jeux de données, ces cas sont exclus de cette étude.

Filtrage sur les visites futures

Le système d'AlayaCare permettant de planifier des visites pour le futur, une certaine quantité de visites sont planifiées à une date postérieure à la date d'extraction des données. Ces visites peuvent être amenées à être modifiées, certaines peuvent être assignées à d'autres employés ou être annulées par le patient. Par ailleurs, l'information exacte du temps de la visite n'est pas encore disponible, il ne s'agit que d'une estimation. Prendre en compte ces visites est une hypothèse trop forte, celles-ci sont retirées de la table VISITES. De cette manière l'information qui est conservée se situe dans le passé uniquement, relativement à la date d'extraction des données.

Système de pointage

À travers l'utilisation de l'application sur téléphone ou à l'aide d'un appel au gestionnaire, les employés peuvent marquer le début et la fin de leurs visites. L'information de la durée exacte de la visite est utilisée par la suite pour les systèmes de paie (employé) et de facturation (patient). Afin de filtrer certaines incohérences, l'hypothèse suivante est proposée afin de vérifier si la durée de pointage est valide :

Hypothèse 1 (H1): *Le pointage d'une visite est considéré valide lorsque la durée de pointage est située entre* ***20%*** *et* ***200%*** *de la durée estimée initialement.*

Cette hypothèse permet d'éliminer les cas où l'employé a pointé le début de la visite, et a oublié de marquer la fin. La fin de la visite peut alors être marquée plusieurs jours après, ce qui fausse l'interprétation de la durée réelle d'une visite. D'autres cas où le pointage de début de visite et de fin de visites sont faits à une intervalle de quelques secondes ont été relevés, introduisant des durées de pointage quasi-nulles. Les visites concernées ne sont pas filtrées, mais les informations de pointage seront marquées comme manquantes et ne seront pas prises en compte.

Il faut préciser que cette fonctionnalité de pointage n'est pas obligatoire, certaines visites n'ont pas cette information. D'autres clients ne l'utilisent pas du tout. L'utilisation peut être appliquée pour des types d'emplois spécifiques seulement, comme les préposés aux services de soutien à la personne et non les infirmières auxiliaires par exemple. Ces informations seront donc utilisées uniquement lorsqu'elles sont disponibles.

Filtrage des durées de visites incohérentes

Des erreurs de saisie de visites peuvent entraîner des durées de visites estimées incohérentes. Afin de filtrer ces quelques cas, l'hypothèse suivante est faite :

Hypothèse 2 (H2): *La durée estimée d'un visite ne peut pas être inférieure à* ***1 minute****, et ne doit pas excéder* ***24 heures****.*

Les visites qui ne font pas partie de cet intervalle sont exclues de la table Visites. Ce choix permet de réduire d'éventuelles erreurs de saisies.

Statut d'approbation et d'annulation

Le statut d'approbation se rapporte à la facturation de la visite au patient. Lors de sa création, une visite est marquée par défaut comme `Non approuvée`. Une fois la visite passée, elle peut être marquée comme `Approuvée` ou `Rejetée` s'il y a eu un problème.

Une autre information disponible est le statut d'annulation d'une visite. Par défaut, une visite est `Maintenue`. De multiples raisons peuvent être la source de l'annulation d'une visite : l'employé peut se destituer, le patient peut ne plus avoir besoin du soins, ou la visite peut être reportée par exemple. Dans ce cas, le statut d'annulation passe à `Annulée`, et un champ de texte permet d'ajouter un commentaire concernant la raison de l'annulation.

Toutes les combinaisons de statuts d'approbation et d'annulation sont envisageables. Dans cette étude, on s'intéresse seulement aux visites qui ont eu lieu. Ainsi les visites suivantes ne seront pas considérées, et seront exclues de la table Visites :

— lorsque le statut d'approbation indique `Rejetée`, quelque soit le statut d'annulation
— lorsque le statut d'annulation indique `Annulée`, quelque soit le statut d'approbation

Cependant, le nombre de visites ayant le statut d'annulation `Annulée` sera relevé par employé comme une variable explicative avant d'être exclues de la table VISITES.

Étiquettes temporelles

En se basant sur la date et l'heure du début d'une visite, on peut extraire plusieurs indices temporels. La première mesure effectuée est de savoir si la visite a lieu pendant la `Semaine` ou pendant la `Fin de semaine`. Une variable booléenne est ajoutée à toutes les visites de la table VISITES lorsque la visite a lieu pendant la `Fin de semaine`.

Une autre variable qui peut être extraite est la période de la journée pendant laquelle la visite est prévue. Trois catégories ont été définies : `Journée`, `Fin de journée` et `Nuit`. L'étiquetage se base seulement sur l'heure de début de la visite, et la classification est faite de la façon suivante en s'inspirant de [23] :

— `Journée` : la visite commence entre 8h et 17h
— `Fin de journée` : la visite commence entre 17h et 23h
— `Nuit` : la visite commence entre 23h et 8h

L'intuition derrière l'extraction de ces variables est de vérifier si la proportion de visites de nuit, en fin de journée ou en fin de semaine a un impact sur le départ des employés comme l'affirme [6, 11, 23]. Pour les visites longues (plusieurs heures) il est possible que la visite soit à cheval entre plusieurs périodes. La durée d'une visite étant généralement d'une heure, ce genre de cas reste minoritaire.

Figure 4.2 Étiquetage temporel des visites en fonction de la période de la journée

Il est aussi possible de relever si une visite a commencé dans les heures de pointe. Les moyens de transport sont généralement saturés, et peuvent être un indicateur de pénibilité. Les heures de pointe dépendent des villes et des pays, mais elles se divisent généralement en deux périodes : le matin autour de 9h lorsque les gens vont au travail, et l'après-midi autour de

16h lorsqu'ils rentrent chez eux. Le découpage utilisé pour cette mesure est présenté sur la figure 4.3. Une variable booléenne `Heures de pointe` est créée pour chaque visite, dont la valeur est `Vraie` lorsque la visite a commencé dans les heures de pointe.

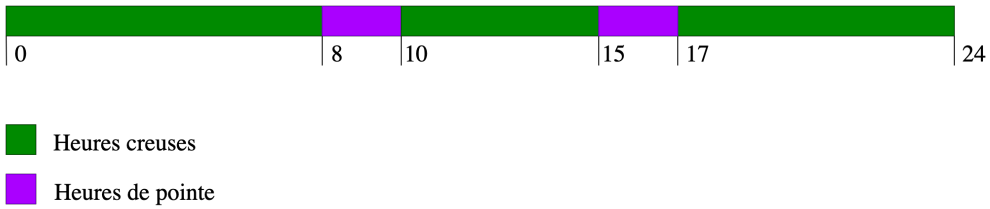

Figure 4.3 Étiquetage temporel des visites en fonction des heures de pointes

Attribution des coordonnées géographiques aux visites

Afin de pouvoir calculer des distances parcourues entre des visites, il faut connaître les coordonnées latitude/longitude de l'endroit où a eu lieu la visite. Dans toutes les tables de données, il y a quatre champs possibles d'adresses :

- Si la visite a eu lieu dans une infrastructure (hôpital, clinique...) alors celle ci sera précisée à l'aide d'un identifiant dont les coordonnées pourront être retrouvées dans la table Infrastructures
- Le créateur de la visite peut spécifier directement dans la visite le lieu de la visite, les coordonnées sont alors disponibles dans la table Services
- Lorsque l'employé effectue son pointage, les coordonnées du pointage peuvent être relevées si l'application mobile est utilisée, l'information se retrouve alors directement dans la table Visites
- les coordonnées du domicile du patient associé à la visite peuvent être accessibles avec la table Patients

Pour une visite donnée, les quatre informations sont rarement toutes disponibles. Pour associer des coordonnées à une visite, il a été choisi une assignation hiérarchique. La première table fouillée sera la table Infrastructures (seulement si une infrastructure a été spécifiée lors de la création de la visite). Si l'information n'est pas disponible, on regarde la table Services. On poursuit de manière analogue avec la table Visites puis la table Patients. Il est possible que l'information ne soit disponible dans aucune des quatre tables susnommées. Dans ce cas, l'adresse de ces visites sera considérée comme manquante. La recherche d'adresse pour une visite s'arrête à la première paire de coordonnées trouvée. Il est important de mentionner que la précision des coordonnées a été légèrement altérée afin de ne pas

pouvoir identifier précisément un lieu de résidence ou un bâtiment spécifique.

4.1.4 Modifications de la table Employés

Processus d'insertion d'une nouvelle entrée dans la base de données

Lorsqu'un employé commence à travailler pour un client, le gestionnaire ou le coordonnateur l'ajoute à sa base de données. Il doit rentrer manuellement plusieurs informations telles que le jour de début de l'activité, la fonction de l'employé, son genre, sa date de naissance ainsi que ses capacités horaires hebdomadaires et journalières (maximales et minimales). Aucun de ces champs n'est obligatoire, il est donc courant que certaines données soient manquantes ou erronées. Par ailleurs, plusieurs champs peuvent être remplis à l'aide d'une saisie de texte manuelle, c'est-à-dire qu'il n'y a pas de menu déroulant avec des options pré-remplies. Une conséquence de cette particularité est que les analyses automatiques des champs textes doivent être robustes aux fautes d'orthographe, à l'alphabet utilisé (accents et caractères spéciaux) ainsi qu'à la présence de lettres majuscules ou minuscules. L'âge des employés est déterminé à partir de leur année de naissance.

Employés sans visites

Après avoir fait le pré-traitement des visites, certains employés n'affichent avoir aucune visite. Ceci peut correspondre à des employés qui viennent de commencer, ou des entrées d'employés administratifs par exemple. Ces employés n'ayant aucune donnée d'activité sur laquelle s'appuyer, ils sont retirés de l'ensemble d'étude. Il est aussi possible que toutes les visites d'un employé aient été filtrées par les critères mentionnés dans la section précédente. Un tel cas signifierait que les données de cet employé ne rentrent pas dans le cadre de cette étude. Par exemple, les données des visites peuvent avoir été mal saisies ou sont absentes, ainsi il est raisonnable de ne pas entraîner les modèles sur ces données.

Nombre de jours d'activité d'un employé

Le nombre de jours d'activité est obtenu de la manière suivante. Pour les employés qui ne sont plus en fonction, on soustrait la date de fin d'activité et celle de début. Pour les employés encore actifs, le nombre de jours d'activité correspond au nombre de jours qu'il s'est écoulé entre la date de début de fonction et la date d'extraction des données du logiciel d'AlayaCare. Cette période de temps est stockée en nombre de jours. Il faut noter que la date de début d'activité n'est pas systématiquement présente. Ceci concerne aussi la date de fin d'activité pour les anciens employés. Si une de ces deux valeurs venait à manquer pour un employé, le

calcul de la durée d'activité n'est pas possible avec cette méthode. Pour des raisons qui sont propres aux utilisateurs du logiciel d'AlayaCare, il est possible que l'employé soit marqué comme terminé longtemps après sa date de départ. Par conséquent il est possible que la date de départ saisie dans la base de données ne soit pas fidèle à la vérité terrain.

Figure 4.4 Temps d'activité d'un employé

Une autre variable contenant le nombre de jours qu'il s'est écoulé entre la première et la dernière visite est calculée. Ces dates sont extraites de la table VISITES et sont ajoutées à la table EMPLOYÉS. L'avantage de cette deuxième mesure de la durée totale d'activité d'un employé est qu'il n'y a pas de données manquantes : à ce stade, tous les employés conservés ont des visites enregistrées, et les dates des visites étant systématiquement remplies, la différence de jours entre les deux visites extrêmes peut être calculée. Au vu des imprécisions de la première mesure, cette donnée d'activité sera privilégiée afin d'apprécier l'ancienneté des employés.

Correction du statut des employés actifs

Lorsqu'un employé est créé, il est mis par défaut sur le statut *Active*. Le statut est susceptible de changer au cours du temps, et peut être mis sur *On-Hold* si l'employé est momentanément indisponible (congé maternité, vacances...), ou sur *Terminated* si l'employé ne travaille plus avec ce client. Lors de l'exploration des données, il a été relevé de nombreux cas où des employés sont marqués *Active* mais n'ont effectué aucune visite dans les mois précédents, voir les années précédentes. En effet, il n'y a aucune obligation pour un gestionnaire de mettre à jour le statut d'un employé lorsqu'il est parti. Ainsi la personne reste *Active* pour une période pendant laquelle il n'est plus en fonction. Afin de corriger ces cas, on propose l'hypothèse suivante :

Hypothèse 3 (H3): *Un employé avec le statut* ***Active*** *qui n'a pas eu de visites pendant les* ***3 derniers mois*** *à compter de la date d'extraction des données est considéré comme étant* ***Terminated****.*

Le facteur 3 mois a été déterminé en se renseignant auprès de clients qui appliquent la politique *90 days no-shifts policy*, soit un employé qui n'a pas montré d'activité dans les trois

derniers mois. Il est possible de créer certaines erreurs dans le cas où l'employé est parti en congé maternité ou en vacances à long terme, mais ces raisons restent minoritaires. Par ailleurs, ces employés devraient avoir le statut *On-Hold*.

Genre des employés

Comme mentionné précédemment, le champ "Genre" rempli lors de la création d'un nouvel employé est un champ texte libre. Les trois catégories utilisées pour classifier le genre utilisés dans cette étude sont : **Homme**, **Femme** ou **Autre**. En prenant en compte que les clients d'AlayaCare écrivent majoritairement en anglais ou en français, l'analyse de ces champs est réalisée en utilisant des expressions régulières. Le genre est inféré à partir des deux champs *Gender* et *Salutation*. Dans le cas où un client n'a rempli aucun de ces deux champs, aucun genre ne peut être inféré à partir des données. Les employés dont les données sont absentes ou ne peuvent pas être déchiffrées sont mis par défaut dans la catégorie **Autre**.

Capacités journalières et hebdomadaires

Lorsqu'un employé est entré dans la base de données, il est possible de préciser :

- une capacité horaire journalière maximale
- une capacité horaire journalière minimale
- une capacité horaire hebdomadaire maximale
- une capacité horaire hebdomadaire minimale

Une fois de plus, ce champ est facultatif et n'est pas systématiquement rempli. Par ailleurs, il existe une valeur qui peut être fixée par défaut par un client pour l'ensemble de ses employés. Le traitement appliqué est le suivant : si la donnée personnalisée n'est pas renseignée, la donnée par défaut est assignée aux capacités horaires journalières et hebdomadaires. Ainsi, si la valeur par défaut est renseignée par un client, il n'y aura pas de données manquantes pour cette variable. En pratique, cette fonctionnalité est peu utilisée. Parmi les quatre variables susnommées, la capacité horaire hebdomadaire maximale est celle qui est la plus remplie. De manière générale, elle est réglée autour de 40 heures par semaine. Les valeurs des trois autres étant presque absentes, on ne conserve que cette variable. Elle permettra par la suite de mesurer si les employés ont fait du temps de travail supplémentaire, au delà du maximum hebdomadaire spécifié, ou si le nombre d'heures donné dans une semaine est anormalement faible [22, 23].

Commentaires lors du départ d'un employé

Lorsqu'un gestionnaire fait passer l'état d'un employé de *Active* vers *Terminated*, il peut ajouter une note expliquant la raison de ce changement. Parmi les causes qui reviennent souvent, on retrouve : les départs à la retraite, les congés maternité, les licenciements ou les départs volontaires par exemple. Le champ associé à ce commentaire étant aussi un champ de texte libre, analyser en profondeur le texte relèverai d'une autre étude. Des expressions régulières ont été appliquées afin de relever les mots se rapprochant des causes les plus probables. L'objectif de notre étude est d'étudier les abandons liés à l'insatisfaction. Il faut donc être en mesure de séparer les causes involontaires comme les licenciements, un congé maternité, ou un départ à la retraite et les causes d'abandon comme un employé qui a démissionné, ou qui n'a plus aucun contact avec l'entreprise. Les trois catégories utilisées pour classifier la raison des départs est la suivante : **Abandon**, **Non-Abandon** et **Inconnue**. Lorsque le commentaire est absent ou ne peut pas être déchiffré, la raison du départ est mise par défaut dans la catégorie **Inconnue**. Dans cette étude, on se focalise sur les personnes qui ont abandonné leur travail, pour des motifs volontaires : on conservera donc uniquement les employés dont le départ a été classifié comme **Abandon**.

Catégories d'emploi

Afin de valider l'hypothèse que le type d'emploi a une influence sur la satisfaction au travail et l'abandon [7, 8, 23, 24], il est nécessaire d'exploiter cette information. La classification utilisée tente de distinguer le personnel médical nécessitant un niveau d'études plus élevé et le reste du personnel de soins. Les catégories utilisées sont : **Personnel médical**, **Autre personnel**, et **Inconnu**. Au vu du nombre de catégories uniques, des expressions régulières sont utilisées afin d'aller chercher les mots nécessaires à la classification. Lorsque la donnée est manquante ou indéchiffrable, l'employé est placé dans la catégorie **Inconnu**. Cette variable explicative catégorielle sera utilisée dans les modèles afin de mesurer son impact sur la satisfaction des employés.

Listes noires

À travers l'interface d'AlayaCare, il est possible pour un employé de bloquer un patient et vice-versa. Une fois la personne mise en liste noire, la personne en charge d'assigner les visites aux employés peuvent éviter les appairages indésirables. Ces données sont stockées dans un fichier à part dans lesquelles on peut relever l'identifiant du bloqueur, du bloqué et la date de création de l'ajout. En regroupant l'information à l'échelle d'un employé, on peut relever :

— le nombre de patients qu'il a bloqué
— le nombre de patients qui l'ont bloqué

L'intuition est qu'un ajout dans une liste noire, que ce soit un patient ou un employé peut être l'indicateur d'un problème, dont les causes peuvent être multiples. Il est possible qu'une dégradation de la satisfaction ou de la santé puisse conduire à plus d'erreurs médicales [22] par exemple, ce qui peut conduire un patient à ne pas être satisfait par la qualité de soins [10].

Il est possible qu'un client n'utilise pas cette fonctionnalité. Ainsi la liste noire sera vide, et les valeurs des deux points cités dans le paragraphe précédent seront nuls pour tous les employés.

Disponibilités des employés

Le logiciel d'AlayaCare permet de rentrer pour chaque employé ses disponibilités ou indisponibilités de la semaine. Il est ainsi possible de savoir quand un employé peut effectuer des visites ou lorsqu'il a une impossibilité. Les congés sont aussi ajoutés à l'emploi du temps. Estimer la disponibilité d'un employé et comprendre la flexibilité de son emploi du temps est un point très important pour les gestionnaires afin de choisir correctement les employés à assigner aux visites. De plus, un employé très disponible qui ne reçoit aucune proposition de visites n'aura pas envie de conserver son emploi. L'ajout de cette fonctionnalité étant très récente dans le système d'AlayaCare, le volume de données est très faible et non systématiquement rempli. Aucune conclusion valable ne pourra donc être tirée de ces informations, mais il serait envisageable de mener l'étude une fois que cette option sera plus largement utilisée.

4.1.5 Méthode de calcul des distances

D'après [8], les distances parcourues par les employés de soins à domicile peut avoir un impact sur leur satisfaction. Ainsi, pour relever cette information, on procède à la méthode décrite ci-après.

Détermination des routes

Dans cette étude, on considère les déplacements entre les lieux des visites et le domicile de l'employé. La nécessité d'avoir un critère de temps pour considérer que deux visites sont adjacentes nous amène à faire l'hypothèse suivante :

Hypothèse 4 (H4): *Si l'intervalle de temps entre deux visites est inférieur à **3 heures**, on considère que l'employé se dirige vers la deuxième visite une fois la première terminée. Dans*

le cas où l'écart est supérieur à cette valeur, alors on considère que l'employé est retourné chez lui à la fin de la visite. Il repartira de son domicile pour aller à la visite suivante.

Une estimation des coordonnées du domicile de l'employé étant généralement fournie, on est capable de dessiner des routes et donc d'estimer la distance parcourue entre les visites, ainsi que la distance parcourue entre le domicile de l'employé et les visites. Le trajet des employés est donc calculé de la manière suivante :

- l'employé part de chez lui et se dirige vers sa première visite
- tant que la visite suivante a lieu moins de 3h après la fin de la visite en cours, l'employé se déplace de visite en visite
- lorsque la visite suivante a lieu plus de 3h après la fin de la visite en cours, l'employé part de sa dernière visite et se dirige vers son domicile

Distance à vol d'oiseau

On peut définir λ_1, ϕ_1 comme les coordonnées longitude et latitude en radians de la première visite. On définit de la même manière λ_2, ϕ_2 pour la seconde. En posant $\Delta\lambda = |\lambda_2 - \lambda_1|$ et $\Delta\phi = |\phi_2 - \phi_1|$, on est capable de connaître l'angle entre les deux points $\Delta\sigma$ à l'aide de la formule de Vicenty :

$$\Delta\sigma = \arctan \frac{\sqrt{\left(\cos\phi_2 \sin(\Delta\lambda)\right)^2 + \left(\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2 \cos(\Delta\lambda)\right)^2}}{\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\Delta\lambda)}$$

Une fois l'écart en angle connu, il suffit de multiplier par le rayon r de la sphère pour obtenir la distance d :

$$d = r\,\Delta\sigma$$

Ainsi, on est capable d'estimer une distance à partir d'une paire de coordonnées. L'ensemble des distances respectant l'hypothèse 4 sont calculées. Dans le cas où des coordonnées étaient manquantes, on extrapole linéairement les distances calculées avec succès.

Une régression linéaire réalisée à partir de requêtes de distances réelles en voiture nous permet de corriger notre calcul de distance et d'avoir une estimation approximative du temps passé dans les transports.

Il n'est pas possible d'avoir accès au réel moyen de transport utilisé par l'employé à chaque déplacement. La voiture étant le moyen le plus commun, il sera considéré comme référence pour les calculs de distances.

4.1.6 Gestion de la superposition de visites

La superposition d'horaires intervient lorsque l'intersection des périodes de deux visites n'est pas nulle. Quatre exemples sont présentés dans la Figure 4.5. Ces quatre exemples concernent une situation où trois visites sont prévues sur un créneau de trois heures, avec des durées et des horaires différentes. Les visites sont triées par heure de début de visite. L'analyse de chaque exemple est la suivante :

- 1e exemple : les trois visites sont superposées entre elles. Elles ont toutes mutuellement une intersection non nulle. La somme des visites planifiées est **6 heures** sur un créneau de 3 heures.
- 2e exemple : seule la visite 2 a une intersection non nulle avec les deux autres. Les visites 1 et 3 sont parfaitement distinctes. La durée totale des visites vaut **4 heures**.
- 3e exemple : comme pour l'exemple précédent, la visite 2 a une intersection non nulle avec la première et la troisième. On peut observer que la visite 3 est intégralement planifiée pendant la visite 2. Cet exemple montre **5 heures** de visites planifiées.
- 4e exemple : les visites 2 et 3 sont intégralement planifiées pendant la première visite. Ainsi la superposition d'horaires ne concerne pas juste les visites immédiatement adjacentes, la source peut être plus lointaine. Pour détecter ce phénomène, il faut donc parcourir l'ensemble du jeu de données pour chaque visite. La somme des visites planifiées donne **4 heures** pour cet exemple.

Pour estimer le temps passé dans chacune des visites lorsque les données de pointage ne sont pas disponibles, on peut calculer la somme de toutes les visites, ou l'écart de temps entre la visite qui commence le plus tôt et celle qui finit le plus tard. La donnée utilisée dépendra du client concerné. On utilisera pour cela les informations des pratiques de celui-ci dont on dispose.

4.1.7 Variables socio-économiques

Au Canada, un code postal est constitué de six caractères. Les trois premiers constituent la Région de Tri d'Acheminement (RTA). Ces régions délimitent une zone géographique dont la taille dépend de la densité d'habitants et de la province concernée. À travers les données de recensement du Canada dont les dernières accessibles datent de 2016 [38], on peut récupérer certaines variables socio-économiques par RTA. Parmi celles-ci, on retrouve :

- nombre d'habitants
- pourcentage de la population mariée
- revenu médian
- revenu moyen

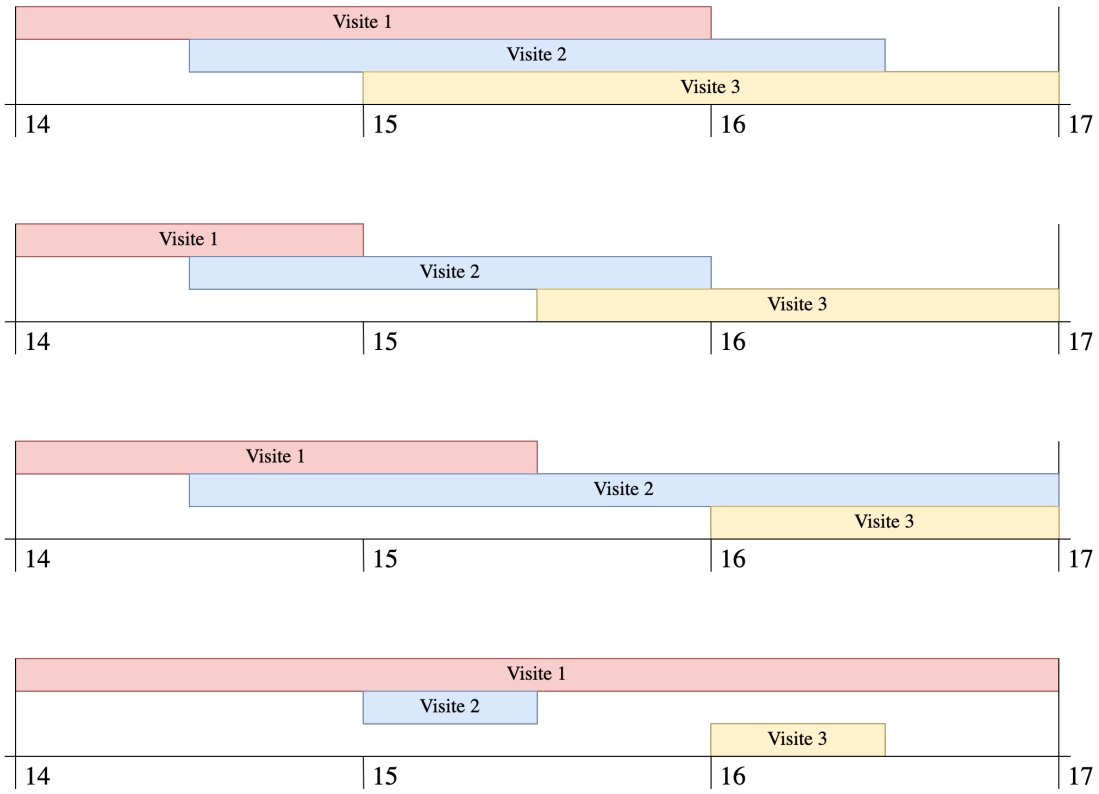

Figure 4.5 Quatre exemples de superpositions d'horaires avec trois visites

— pourcentage de la population dont la langue maternelle est une langue officielle locale
— pourcentage de citoyens canadiens parmi la population
— nombre de personnes avec un diplôme
— nombre de personnes avec un diplôme du baccalauréat ou supérieur
— pourcentage d'utilisation de la voiture comme mode de transport principal
— pourcentage d'utilisation des transports en communs comme mode de transport principal

À partir des RTA du domicile des employés, on est capable d'avoir des indicateurs sur le milieu de vie de celui-ci. Les informations extraites du recensement peuvent apporter de l'information sur des variables identifiées par la littérature dans les sections 2.2.1 et 2.2.3 : ethnicité, langue parlée, niveaux de salaire, types de fonction qui dépend entre autres des études réalisées, la distance parcourue, ou encore la ruralité d'une agence de soins.

On peut aussi relever ces variables à partir du RTA des patients. À l'échelle d'un employé, on peut relever les moyennes de ces variables pour les patients vus pendant une période afin d'avoir une idée du milieu de travail dans lequel celui-ci travaille. Ce type de données pourra

donc être utilisé pour les clients canadiens.

Pour les États-Unis, nos recherches ont permis d'avoir accès à la population par code postal, mais pas d'autres variables socio-économiques. Ces variables ont été ajoutées pour les clients états-uniens. Pour l'Australie, en plus de la population par code postal, le gouvernement a mis à disposition l'IRSAD (*Index of Relative Socio-economic Advantage and Disadvantage*) qui prend en compte les avantages et inconvénients en lien avec l'emploi et les revenus sur les différents codes postaux. Le tableau A.2 résume les variables socio-économiques inclues dans cette étude selon le pays.

4.2 Préparation des données

4.2.1 Critère de sélection des clients

Les clients d'AlayaCare ayant tous des besoins différents, la plateforme n'est pas utilisée de la même manière par tous. Les conséquences directes sont que certaines variables explicatives peuvent ne pas être présentes dans les données de certains clients. De plus, il faut que les modèles soient entraînés avec suffisamment de données. Les filtres suivant ont été appliqué à l'échelle des clients afin d'identifier les jeux de données avec lesquels travailler :

- Volume de visites enregistrées : supérieur à **20000**
- Intervalle de temps entre la première visite enregistrée et la date d'extraction des données : supérieur à **6 mois**
- Nombre d'employés totaux : supérieur à **500**
- Nombre d'employés encore actifs : supérieur à **100**
- Nombre d'employés qui ne sont plus en fonction : supérieur à **100**
- Nombre de visites en superposition d'horaire : moins de **20%** des visites totales
- Coordonnées des lieux de visites : présentes à plus de **90%** afin de développer une mesure de distances parcourues

Cette analyse préliminaire nous permet d'identifier 10 clients potentiels. Ceux-ci seront nommés Company_1, Company_2, jusqu'à Company_10 afin de conserver leur anonymat.

4.2.2 Création des jeux d'entraînement et étiquetage

Afin de pouvoir utiliser les algorithmes de classification tels que les régressions logistiques, les arbres de décision et les réseaux de neurones, il faut construire une base de données d'entraînement. Pour notre problématique, la classification ne peut pas être réalisée sur les employés directement car la variable de satisfaction évolue au cours du temps. La stratégie proposée est de considérer des périodes d'activités d'un employé. En effet, un employé peut

avoir des périodes plus ou moins stables avant de prendre la décision de quitter son emploi. De plus, le statut des employés encore en fonction est susceptible d'évoluer dans les semaines suivant la date d'extraction des données. Les modèles s'appuieront donc sur les données des employés déjà partis. Les modèles ainsi créés seront ensuite appliqués sur les employés encore actifs dans l'objectif de mesurer le taux de stabilité global des employés d'un client (méthode décrite à la section 5.7).

Étiquetage des périodes d'activité des employés terminés

Les instances de la base de données d'entraînement que l'on souhaite créer sont des périodes d'activité d'un employé. Les variables explicatives seront donc relevées pour chaque période. Pour l'étiquetage, on fait l'hypothèse suivante :

Hypothèse 5: *Les périodes proches de la date de départ d'un employé sont associées à un risque plus élevé d'instabilité que les périodes plus éloignées dans le temps.*

Ainsi, la situation d'un employé au début de son emploi est considérée comme stable, contrairement aux périodes précédant son départ. L'objectif des algorithmes de classification est de distinguer les périodes de fin d'emploi parmi toutes les périodes. Le problème est abordé comme un problème de classification binaire, de la même manière que [23, 26]. Les deux classes considérées sont Période Stable et Période à Risque. Plusieurs paramètres doivent être déterminés, un exemple est proposé sur la figure 4.6 :

- durée d'une période : en nombre de jours, la durée d'une période dans l'ensemble d'entraînement quelle que soit son étiquette
- nombre de périodes Période à Risque : nombre de périodes à relever pour les employés qui ne sont plus en fonction. Ces périodes constituent les instances de la classe positive
- nombre de périodes frontières à ignorer : afin de mieux distinguer les deux classes citées précédemment, les données d'un certain nombre de périodes après les périodes étiquetées comme Période à Risque seront mises de côté. Ce choix doit permettre aux algorithmes de classification de mieux distinguer les écarts entre les deux classes. Les données de ces périodes ne seront pas utilisées
- nombre de périodes Période Stable : ces périodes doivent se retrouver suffisamment loin de la date d'extraction pour être considérées comme stable. Ces périodes constitueront les exemples de la classe négative

Pour l'exemple de la figure 4.6, les paramètres utilisés sont les suivants :

- durée d'une période : **7 jours**

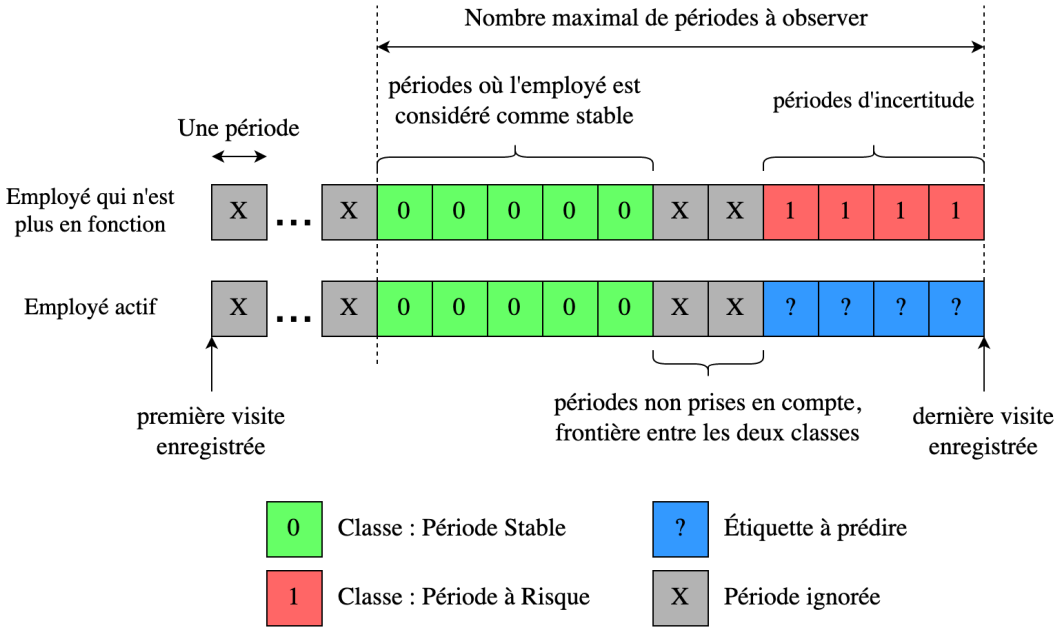

Figure 4.6 Exemple d'étiquetage des périodes d'un employé

— nombre de périodes Période à Risque : **4**
— nombre de périodes frontières à ignorer : **2**
— nombre de périodes Période Stable : **5**

À partir des trois derniers paramètres, on peut déduire le nombre total de périodes observées : **11**, dont 2 sont ignorées. Les périodes antérieures à la première visite d'un employé ne sont pas considérées. Il est possible qu'un employé n'ait pas autant de périodes d'activité que le nombre maximal de périodes à observer s'il vient de commencer ou que son contrat n'a duré que quelques semaines : dans tous les cas, l'étiquetage commence à partir de la dernière visite. Si aucune visite n'a été observée sur une période à étiqueter, toutes les données d'activité sont nulles. Les variables explicatives des employés seront tout de même présentes. Si l'employé était en vacances sur une période, celle-ci n'est pas prise en compte : l'employé n'était pas censé travailler. Les différents paramètres essayés dans cette étude sont présentés à la section 4.2.3.

Étiquetage des périodes d'activité des employés encore en fonction

La date de départ des employés encore actifs étant encore inconnue, on ne peut pas savoir si les périodes les plus récentes sont stables ou à risque. Ces périodes ne sont donc pas utilisées dans l'ensemble d'entraînement. Cependant, les données présentes dans le passé

des employés actifs peuvent être utilisées pour augmenter le nombre d'exemples de la classe Période Stable. Celles-ci doivent être prises suffisamment loin des dernières semaines afin de ne pas être proche d'une hypothétique date de départ.

Pour les périodes les plus récentes, une prédiction pourra être réalisée avec les modèles de classification déjà entraînés afin d'estimer la tendance générale de la stabilité de l'ensemble des employés encore en fonction. L'objectif n'est pas de prédire à un niveau individuel, mais bien à l'ensemble des employés d'une entreprise. L'interprétation et le traitement réalisé avec les prédictions de l'étiquette des dernières périodes sont présentés dans la section 5.7.

Influence des paramètres sur l'équilibre des classes

Les paramètres présentés précédemment ont tous une influence dans la quantité de périodes qui seront prises en compte dans chaque classe. On identifie les sources de déséquilibre suivantes :

- selon les clients, la proportion d'employés actifs par rapport au nombre total d'employés enregistrés dans le système d'AlayaCare peut varier. Comme ceux-ci n'apportent aucune instance à la classe positive, une proportion d'employés actifs élevée augmente la taille de la classe négative.
- On peut augmenter le nombre d'exemples de la classe positive en augmentant le nombre de périodes à étiqueter pour les dernières périodes d'un employé qui n'est plus actif. La valeur de ce paramètre doit être choisi afin de refléter au mieux ce qui se passe dans la réalité et non pour simplement augmenter le nombre d'exemples de cette classe. La réflexion est analogue pour la classe négative.
- Si l'historique d'un employé parti est trop faible, les seules périodes étiquetées appartiendront à la classe positive car elles sont toutes proche du départ de l'employé. En effet, l'étiquetage commence à partir de la dernière visite.

4.2.3 Paramètres de découpage testés dans cette étude

Il est nécessaire de valider différentes valeurs de taille de période afin de trouver celle qui correspond le plus à notre problème. Des périodes de 7 jours, 14 jours et 28 jours seront testées. En imposant que le nombre de périodes Période à Risque, de périodes frontières et de périodes Période Stable soient les mêmes, il ne reste que deux paramètres a, b à fixer :

- a : le nombre de semaines dans une période
- b : l'horizon temporel total en semaines

Ainsi, un découpage peut être représenté avec deux entiers a et b sous la forme a/b. Cette re-

présentation sera utilisée pour présenter les résultats du chapitre 5. Les différents paramètres essayés sont présentés dans la table 4.2. Une visualisation graphique des découpages réalisés est présentée sur la figure 4.7.

Tableau 4.2 Paramètres de découpage explorés

Jours dans une période	Période à Risque	Périodes frontières	Période Stable	Représentation a/b
7	2	2	2	1/6
14	1	1	1	2/6
7	4	4	4	1/12
14	2	2	2	2/12
28	1	1	1	4/12
7	8	8	8	1/24
14	4	4	4	2/24
28	2	2	2	4/24

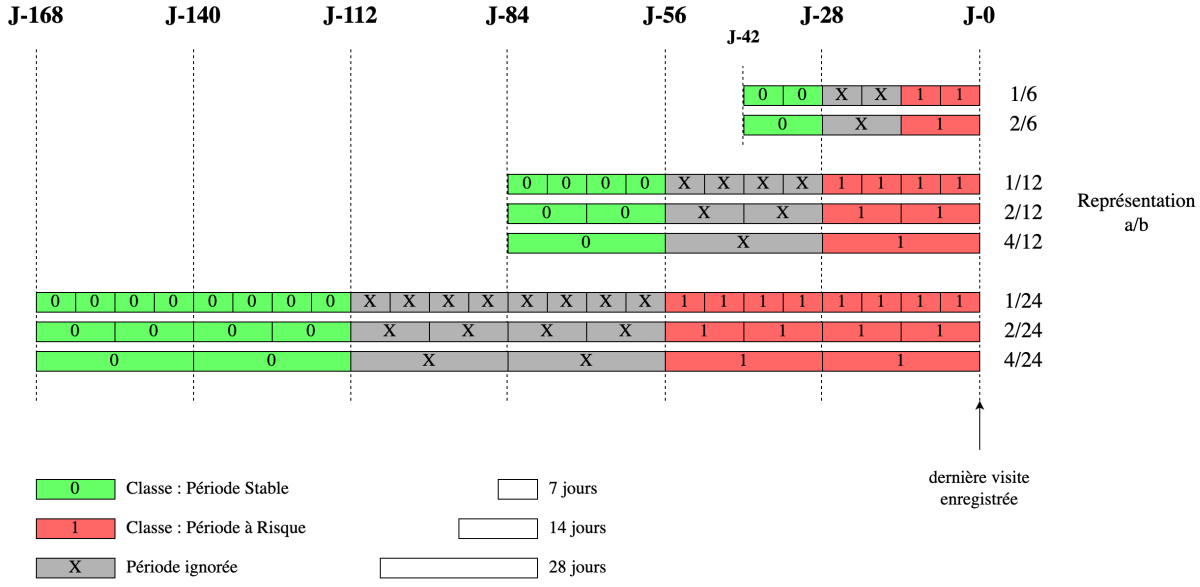

Figure 4.7 Paramètres de découpage explorés pour les employés terminés

4.2.4 Agrégation des données par employés

Méthode

Afin d'obtenir des variables explicatives pour une période, il faut collecter toutes les visites sur la période concernée. Les fonctions d'agrégations sont appliquées à l'ensemble des visites d'une période afin d'en extraire des valeurs numériques pour chaque variable. Les fonctions utilisées peuvent être différentes selon l'information que l'on cherche à obtenir :

- somme : nombre d'heures travaillées, distances parcourues...
- nombre d'entrées : nombre de visites totales, nombre de visites à une certaine période de la journée...
- nombre d'entrées distinctes : nombre de patients différents vus, nombre de jours travaillés...
- moyenne : durée moyenne des visites sur une période
- écart-type : dispersion des durées des visites sur une période

Variables conservées et transformées

Lors de l'agrégation des visites par employés les informations présentes dans le tableau 4.3 sont calculées. En utilisant le coefficient de corrélation présentés au chapitre 3, on remarque que certaines variables comme le nombre d'heures travaillées et le nombre de visites sont très corrélées entre elles. Ainsi, il a été choisi de faire des moyennes : on relève donc la durée moyenne des visites plutôt que la somme des durées des visites. Enfin, les variables qui étaient présentes dans la table Employés suivantes sont ajoutées aux jeux de données ainsi crées :

- `est_homme` : si le `gender` a pu être identifié comme appartenant à la classe **Homme**.
- `est_médical` : variable binaire si le métier d'un employé a été identifié comme **Personnel médical**
- `age` : l'âge de l'employé

Le tableau final des variables utilisées est présenté dans le tableau A.1.

4.3 Modélisation

4.3.1 Métriques d'évaluation

Les modèles utilisés dans cette étude sont des classificateurs. Leurs performances seront évaluées en s'intéressant à l'aire sous la courbe ROC, ainsi que la F_1 de la classe positive comme présenté à la section 3.3. Seuls les résultats de l'ensemble de test seront utilisés pour mesurer les performances d'un modèle. Comparer les résultats entre l'ensemble d'entraînement et de

Tableau 4.3 Variables agrégées par période pour chaque employé

Variable	Signification	Méthode
n_visites	Nombre de visites	somme
heures_par_visite	Durée moyenne d'une visite	moyenne
heures_std	Écart-type des durées de visites	écart-type
retard_par_visite	Retard moyen aux visites	moyenne
n_clients	Nombre de clients distincts	uniques
n_services	Nombre de services différents effectués	uniques
n_annulations	Nombre de visites annulées	somme
distance_par_jour	Distance moyenne parcourue par jour travaillé	moyenne
perc_weekend	Part de visites en fin de semaine	moyenne
perc_nuit	Part de visites de nuit	moyenne
perc_soir	Part de visites en soirée	moyenne
perc_h_pointe	Part de visites en heure de pointe	moyenne
moy_jours_entre_shifts	Nombre de jours moyens entre deux *shifts*	moyenne
n_liste_noire	Nombre d'entrées dans la liste noire	somme
heures_dispo	Nombre d'heures de disponibilité	somme
heures_non_dispo	Nombre d'heures d'indisponibilité	somme

test sera utile pour mesurer les effets du sur-apprentissage. Pour les modèles d'arbres, la probabilité d'appartenir à la classe positive peut être obtenue en considérant la distribution des classes à chaque feuille. On sera alors en mesure de tracer une courbe ROC [33]. Pour les réseaux de neurones, lorsqu'une fonction sigmoïde est utilisée, il suffit de récupérer la valeur du neurone de sortie après la fonction d'activation.

4.3.2 Méthodes de recherche dans un espace de paramètres

Afin de déterminer le jeu d'hyper-paramètres optimaux, on utilise deux outils : la recherche par grille, et la recherche aléatoire. Ces deux méthodes sont utilisées avec une procédure de validation simple afin de comparer les performances et sélectionner les paramètres optimaux.

Recherche par grille

Pour la recherche par grille, on spécifie un ensemble de valeurs que peut prendre chaque hyper-paramètre, et toutes les combinaisons sont testées. On conserve les paramètres du modèle qui a obtenu les meilleures performances sur l'ensemble de validation.

Recherche aléatoire

Contrairement à la recherche par grille, la recherche aléatoire va rechercher un sous-ensemble de paramètres dans un espace défini. Ainsi, on précise une distribution de probabilités pour chaque hyperparamètre, et un nombre d'itération maximal. À chaque itération, chaque paramètre est tiré aléatoirement, et les performances du modèle sont relevées. Le modèle qui obtient les meilleures performances sur l'ensemble de validation est conservé.

4.3.3 Méthodes d'entraînement

Tous les exemples utilisés pour les modèles sont séparés en trois ensembles [37] :

- l'ensemble d'entraînement : permet au modèle d'ajuster ses paramètres afin de s'adapter aux données.
- l'ensemble de validation : ensemble d'exemples non-utilisés pendant l'entraînement qui permet d'évaluer les performances d'un modèle. Cet ensemble est utilisé afin de sélectionner les meilleurs hyper-paramètres
- l'ensemble de test : ensemble d'exemples non utilisés pendant l'entraînement qui permet de mesurer les performances d'un modèle une fois les hyper-paramètres sélectionnés

Les proportions d'exemples affectés à chaque ensemble sont les suivantes : 50% pour l'ensemble d'entraînement, 20% pour l'ensemble de validation, et 30% pour l'ensemble de test. La séparation est stratifiée, cela veut dire que les classes conservent la même proportion d'exemples dans les trois ensembles. De plus, la séparation d'exemples dans les ensembles est reproduite à l'identique pour tous les essais d'une compagnie : pour chaque type de modèle, et chaque type de découpage. Ainsi, les résultats sont comparables puisque les mêmes données auront été vues à travers les différentes expériences réalisées.

4.3.4 Gestion de l'équilibre des classes

Plusieurs méthodes sont envisagées pour équilibrer les classes dans l'ensemble d'entraînement. Elles se distinguent en deux catégories : les méthodes de sur-échantillonnage et de sous-échantillonnage.

Stratégies de sur-échantillonnage

SMOTE [34] est une des méthodes de sur-échantillonnage utilisée. En créant des observations synthétiques, on est capable de balancer les classes jusqu'à un rapport de 1. On utilisera la fonction `SMOTE` de la librairie *imblearn*, dont le paramètre `sampling_strategy`

permet de déterminer la proportion cible entre la classe positive et négative. Les mêmes paramètres sont disponibles pour la méthode de sur-échantillonnage aléatoire avec la fonction `RandomOverSampler` de la même librairie.

Stratégies de sous-échantillonnage

Pour le sous-échantillonnage aléatoire, on utilisera la fonction `RandomUnderSampler`. Cette fonction partage aussi le paramètre `sampling_strategy` qui permet de déterminer l'équilibre des classes désiré. Le sous-échantillonnage avec liens Tomek [36] et la méthode *Edited Nearest Neighbors* sont disponibles avec les fonctions `TomekLinks` et `EditedNearestNeighbours` respectivement. Ces méthodes ne permettent pas d'atteindre une proportion de classe cible mais suppriment des instances avec des considérations de distance.

Combinaisons de stratégies de ré-échantillonnage

Certaines méthodes de sur-échantillonnage et de sous-échantillonnage peuvent être combinées afin de déterminer la meilleure stratégie. Les stratégies couvertes dans cette étude sont inspirées de [33] :

- Ensemble de base non-équilibré (BASE)
- *Random Over-Sampling* (ROS)
- *Random Under-Sampling* (RUS)
- *SMOTE* (SMT)
- *SMOTE* + Liens Tomek (SMT+TMK)
- *SMOTE* + *Edited Nearest Neighbors* (SMT+ENN)

Afin de déterminer la valeur optimale du paramètre `sampling_strategy` des méthodes *Random Under-Sampling*, *Random Over-Sampling* et *SMOTE*, différentes valeurs sont testées pour chaque compagnie et chaque découpage sur un modèle de base. La valeur du paramètre qui donne les meilleures performances est conservée pour chaque découpage et chaque compagnie.

4.3.5 Modèles utilisés et hyperparamètres

Dans cette étude, quatre types de modèles sont explorés : régression logistique, forêts aléatoires, *XGBoost*, et réseaux de neurones. Les fonctions et les librairies utilisées sont les suivantes :

- `linear_model.LogisticRegression` de la librairie *sklearn*
- `ensemble.RandomForestClassifier` de la librairie *sklearn*

— `XGBClassifier` de la librairie *xgboost*, adaptée pour fonctionner avec la librairie *sklearn*
— `neural_network.MLPClassifier` de la librairie *sklearn*

Les hyperparamètres de chaque modèle sont présentés dans les sections suivantes. Les méthodes de recherche d'hyper-paramètres dans un espace ont été présentées à la section 4.3.2.

Modèles de régression logistique

Pour les modèles de régression logistique, les hyperparamètres que l'on cherche à déterminer sont les suivants :

— `l1_ratio` : Pourcentage entre 0 et 1 de l'importance du terme de régularisation de norme 1 par rapport au terme de norme 2. La valeur de 1 correspond à une pénalité L_1 seulement. La valeur de 0 est associé avec une pénalité L_2 seulement. Les valeurs intermédiaires correspondent au modèle *elasticnet* qui ajoute une pénalité L_1 et L_2 avec différents coefficients.
— `C`, la force de régularisation : Impact de la régularisation par rapport à la fonction de perte, des valeurs élevées sont associées à une force plus faible

Modèles de forêts aléatoires

Les hyperparamètres qui doivent être déterminés pour les modèles de forêts aléatoires sont les suivants :

— `n_estimators` : le nombre d'arbres dans la forêt
— `max_depth` : la profondeur maximale de chaque arbre
— `min_samples_split` : le nombre minimal d'exemples pour autoriser une séparation dans un noeud interne
— `min_samples_leaf` : le nombre minimal d'exemples autorisés sur les feuilles de l'arbre après une séparation
— `max_features` : le nombre de variables qui peuvent être parcourues avant la séparation d'un noeud interne

Modèle *XGBoost*

L'architecture *XGBoost* étant aussi un modèle d'arbres, on retrouve des paramètres semblables aux forêts aléatoires. Cependant, la méthode et la librairie utilisée étant différentes, le nom des paramètres peuvent être différents. Les hyperparamètres que l'on cherche à déterminer sont :

— `learning_rate` : taux d'apprentissage
— `subsample` : proportion de l'échantillon d'entraînement utilisé à chaque étape afin de limiter le sur-apprentissage
— `max_depth` : même paramètre que pour les forêts aléatoires
— `colsample_bytree` : idem que `max_features` pour les forêts aléatoires
— `min_child_weight` : idem que `min_samples_split` pour les forêts aléatoires

Réseaux de neurones

Les réseaux de neurones utilisés suivent une architecture *Multi Layer Perceptron*. Les hyperparamètres à déterminer sont les suivants :

— `hidden_layer_sizes` : taille et nombre de couches cachées
— `activation` : fonction d'activation appliquée en sortie des neurones
— `solver` : solveur utilisé pour l'optimisation
— `alpha` : importance de la pénalité L_2
— `learning_rate` : taux d'apprentissage

CHAPITRE 5 RÉSULTATS ET DISCUSSION

La méthodologie appliquée et le détail des expériences ont été présentées dans les chapitres précédents. Cette partie est consacrée à la présentation des résultats. Les figures sont d'abord présentées et expliquées. Tous les commentaires et toutes les analyses et discussions sont présentés à la section 5.5.

Au total, les résultats regroupent :

— 10 clients différents
— 4 types de modèles
— 8 découpages différents

On dénombre donc 320 modèles retenus. Explorer les résultats pour chacun d'entre eux n'est pas envisageable, les figures de cette section présentent les meilleurs modèles ou des distributions de résultats.

5.1 Stratégie d'équilibre des classes

Les différentes stratégies envisagées pour équilibrer les classes ont été présentées à la section 4.3.4. Les métriques observées sont la $F_{1,test}$ de la classe Période à Risque ainsi que l'aire sous la courbe ROC (ROCAUC). La figure 5.1 montre les méthodes d'équilibrage qui ont donné les meilleurs résultats. Cet histogramme a été crée en prenant les meilleurs modèles sélectionnés sur la $F_{1,test}$ de chaque étiquetage (8), et de chaque compagnie (10), soit 80 modèles au total.

5.2 Différentes méthodes d'étiquetage

La figure 5.2 est réalisée en prenant les meilleurs modèles sélectionnés sur la $F_{1,test}$ de la classe positive pour chaque étiquetage et chaque compagnie. Les scores affichés sont respectivement la F_1 et la ROCAUC de l'ensemble de test. L'axe des abscisses représente l'étiquetage sous la forme a/b comme présenté à la section 4.2.3 : a représente le nombre de semaines dans une période et b l'horizon de données observées pour chaque employé en semaines. Les découpages réalisés ont été présentés sur la figure 4.7.

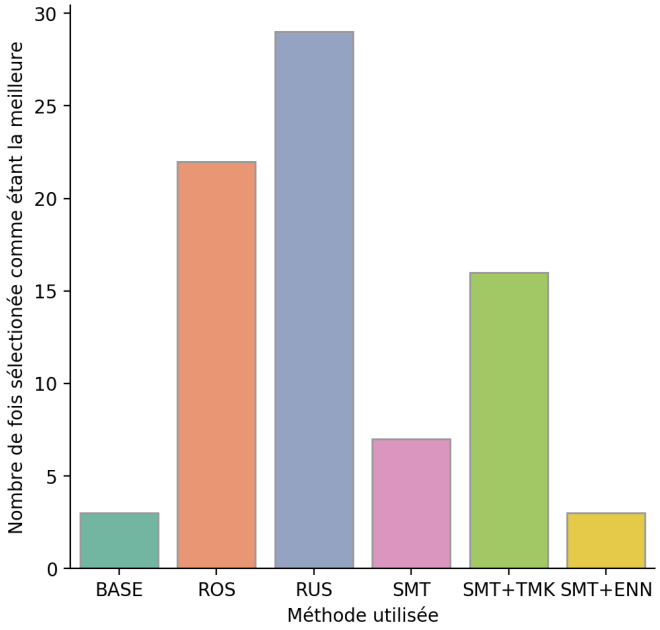

Figure 5.1 Histogramme des méthodes d'équilibrage des classes sélectionnées

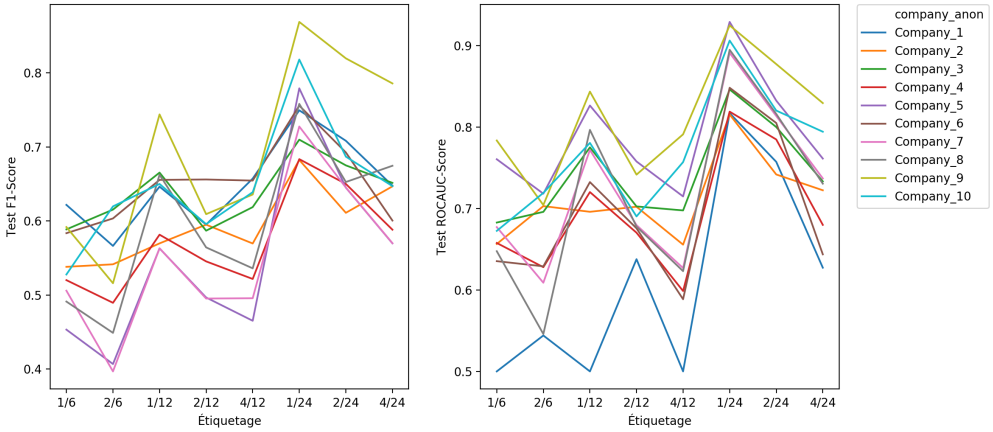

Figure 5.2 Performance en fonction des compagnies et de l'étiquetage

5.3 Performances des modèles

La figure 5.3 montre la $F_{1,test}$ du meilleur modèle pour chaque étiquetage et pour chaque compagnie. La couleur de chaque point indique l'étiquetage utilisé. En abscisse est représenté la proportion d'exemples de la classe positive par rapport au nombre total d'exemples avant les stratégies d'échantillonnage. La ligne en pointillés rouges *dummy* montre les performances d'un algorithme naïf qui prédit toujours la classe positive comme présenté à la section 3.3.

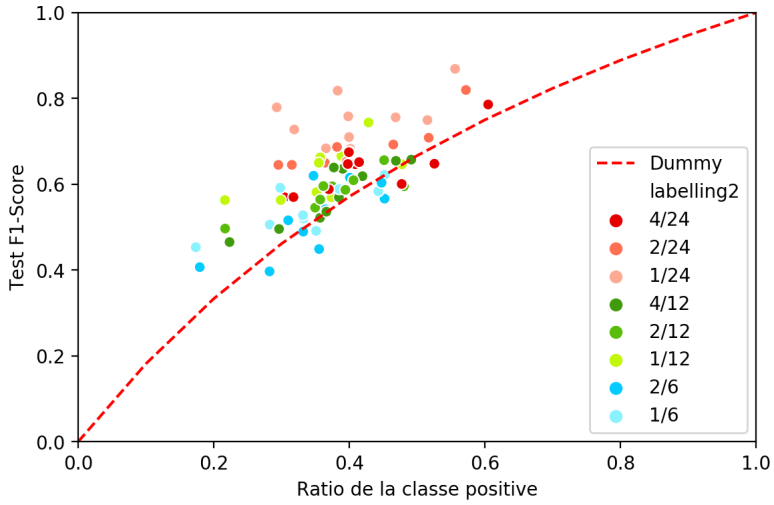

Figure 5.3 $F_{1,test}$ en fonction de la proportion de la classe positive

La figure 5.4 présente la répartition des performances (ROCAUC-Test) des différents types de modèles par client en fonction de l'étiquetage utilisé. De manière analogue, la figure 5.5 montre les performances des différents types de modèles par étiquetages en fonction des clients.

Afin de mesurer le sur-apprentissage, on relève la répartition des rapports de performances entre l'ensemble d'entraînement et l'ensemble de test. Sur la figure 5.6, on représente la valeur suivante :

$$\text{Rapport} = 100 \cdot \frac{F_{1,entraînement} - F_{1,test}}{F_{1,entraînement}}$$

Ce rapport est d'autant plus élevé si les performances de l'ensemble de test sont très différentes de celles de l'ensemble d'entraînement, ce qui est caractéristique du sur-apprentissage. À l'inverse, un rapport de 0 indique que les performances sur ces deux ensembles sont égales.

Le tableau 5.1 montre les performances du modèle qui a le mieux performé pour chaque client, avec le type de modèle associé, l'étiquetage utilisé, et la méthode d'équilibrage de classes.

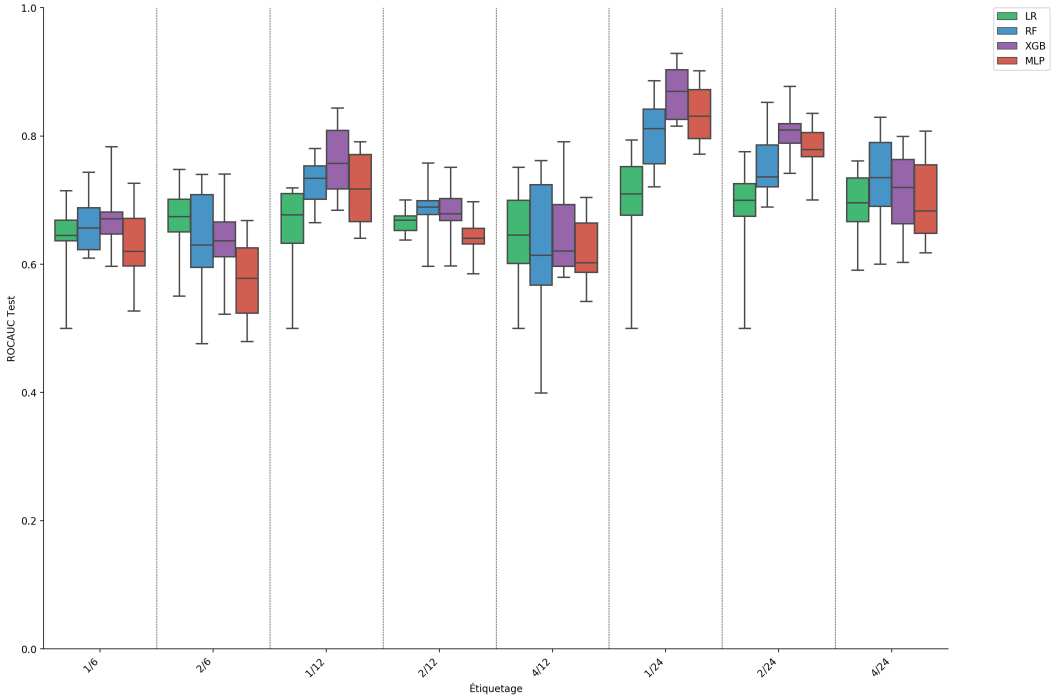

Figure 5.4 ROCAUC des différents types de modèles sur les différents étiquetages

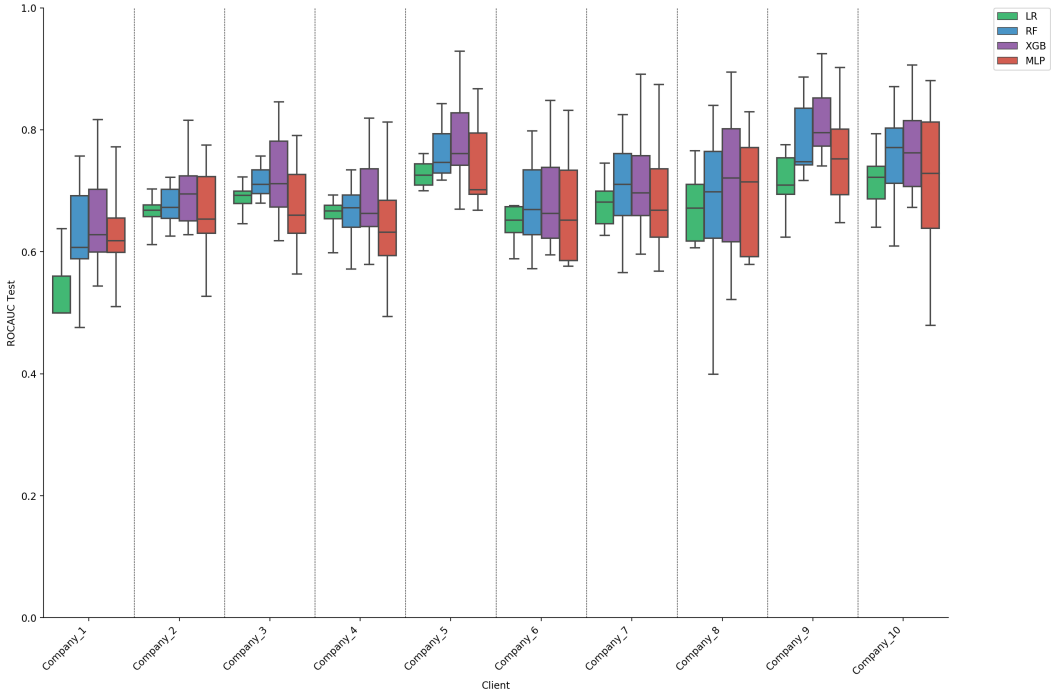

Figure 5.5 ROCAUC des différents types de modèles sur les différents clients

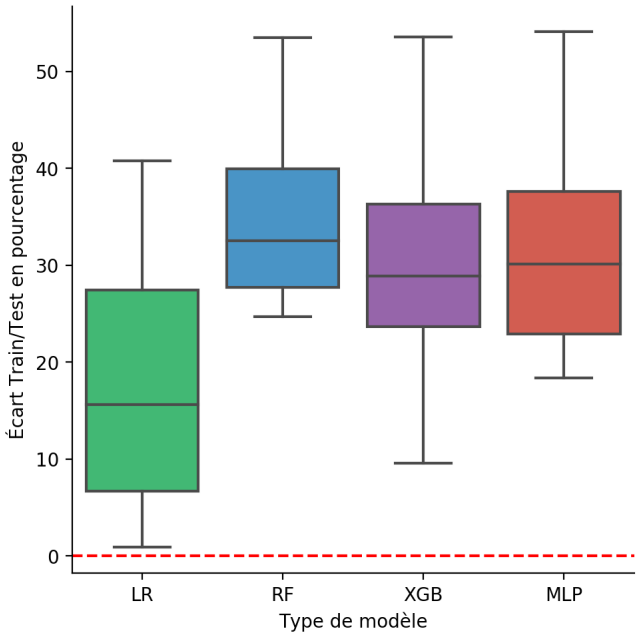

Figure 5.6 Écart entre $F_{1,entraînement}$ et $F_{1,test}$

Tableau 5.1 Performances des meilleurs modèles et étiquetage pour chaque compagnie

Client	Étiquetage	Pays	Modèle	Équilibre	ROCAUC Train	ROCAUC Test	F1 Train	F1 Test	F1 Naïf	Signal (%)
Company_1	1/24	CAN	XGB	ROS	0.991	0.817	0.953	0.749	0.680	21.7
Company_2	1/24	CAN	XGB	ROS	0.975	0.815	0.917	0.682	0.573	25.6
Company_3	1/24	USA	XGB	ROS	0.972	0.846	0.922	0.710	0.571	32.4
Company_4	1/24	CAN	XGB	RUS	0.963	0.819	0.914	0.683	0.535	31.9
Company_5	1/24	AUS	XGB	ROS	0.993	0.929	0.976	0.779	0.453	59.6
Company_6	1/24	CAN	XGB	ROS	0.972	0.848	0.920	0.756	0.638	32.6
Company_7	1/24	AUS	XGB	SMT	0.976	0.891	0.939	0.728	0.483	47.2
Company_8	1/24	USA	XGB	RUS	0.977	0.895	0.930	0.758	0.570	43.8
Company_9	1/24	CAN	XGB	BASE	0.992	0.925	0.961	0.869	0.715	54.0
Company_10	1/24	CAN	XGB	SMT	0.992	0.906	0.957	0.818	0.554	59.2

5.4 Influence des variables explicatives

Le tableau 5.2 présente les moyennes des coefficients obtenus pour les différentes compagnies et les différents découpages avec le modèle de régression logistique. Pour des raisons de visibilité et de clarté, seules les 9 variables avec les plus grandes moyennes en valeur absolue on été conservées. Les tableaux 5.3 et 5.4 présentent les importances des variables en pourcentage pour les modèles de forêts aléatoires et XGBoost respectivement. Dans les deux cas, seules les 9 variables les plus significatives ont été présentées. L'objectif de ces trois tableaux est de visualiser les variables qui ont eu de l'importance lors de la classification.

La figure 5.7 montre les coefficients obtenus avec le modèle régression logistique qui a obtenu les meilleures performances pour chaque compagnie.

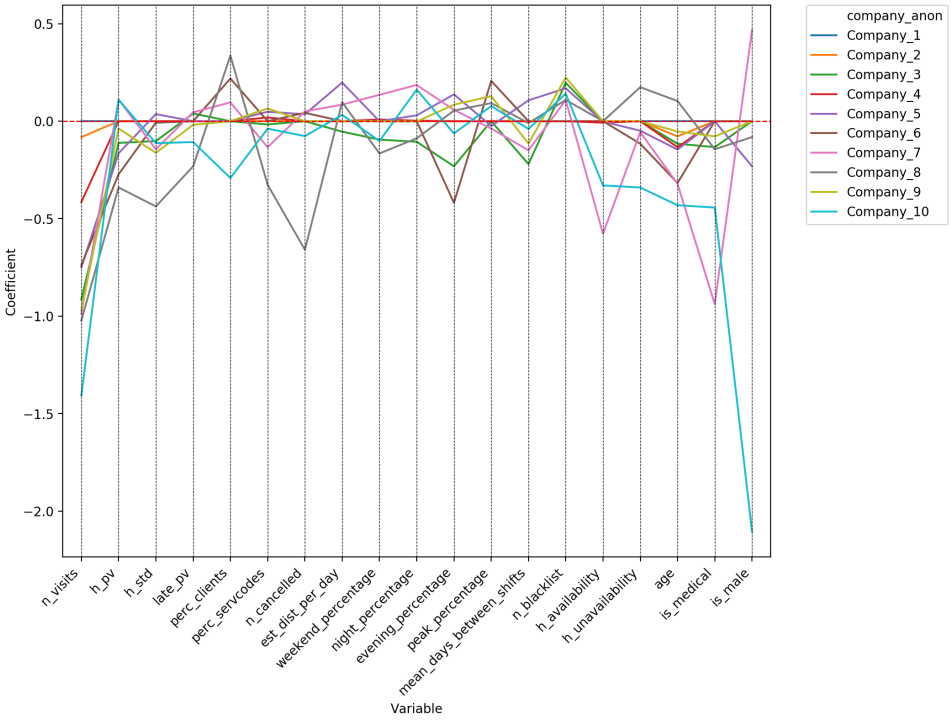

Figure 5.7 Coefficients des variables communes entre tous les pays (Régression Logistique) selon chaque compagnie pour l'étiquetage 1/24

Tableau 5.2 Moyenne des coefficients significatifs des régressions logistiques pour toutes les compagnies et étiquetages

Variable	Moyenne	Écart type	CI 95% Haut	CI 95% Bas
cli_perc_married	0.101	0.267	0.177	0.026
emp_perc_transp_car	0.088	0.272	0.165	0.011
h_pv	-0.090	0.183	-0.050	-0.130
h_unavailability	-0.099	0.375	-0.016	-0.181
is_male	-0.106	0.553	0.015	-0.228
cli_perc_citizen_can	-0.124	0.444	0.002	-0.249
h_std	-0.168	0.342	-0.094	-0.243
age	-0.172	0.198	-0.129	-0.215
n_visits	-0.749	0.722	-0.590	-0.907

Tableau 5.3 Moyenne des importances des coefficients des forêts aléatoires pour toutes les compagnies et étiquetages

Variable	Moyenne	Écart type	CI 95% Haut	CI 95% Bas
n_visits	0.118	0.065	0.132	0.104
perc_servcodes	0.065	0.051	0.076	0.053
perc_clients	0.054	0.032	0.061	0.047
est_dist_per_day	0.049	0.033	0.056	0.042
age	0.048	0.027	0.054	0.042
h_pv	0.048	0.031	0.054	0.041
mean_days_between_shifts	0.045	0.036	0.053	0.037
late_pv	0.045	0.029	0.051	0.038
cli_irsad	0.043	0.013	0.050	0.037

Tableau 5.4 Moyenne des importances des coefficients des modèles XGBoost pour toutes les compagnies et étiquetages

Variable	Moyenne	Écart type	CI 95% Haut	CI 95% Bas
n_visits	0.071	0.058	0.084	0.058
est_dist_per_day	0.070	0.044	0.080	0.061
age	0.066	0.042	0.075	0.057
late_pv	0.064	0.037	0.072	0.056
h_pv	0.062	0.038	0.070	0.054
emp_irsad	0.061	0.027	0.074	0.048
cli_irsad	0.061	0.028	0.074	0.047
h_std	0.047	0.036	0.055	0.039
perc_clients	0.046	0.066	0.060	0.031

5.5 Discussion des résultats

5.5.1 Stratégie d'équilibre des classes

D'après la figure 5.1, on observe que les stratégies de ré-échantillonnage ont permis d'améliorer la classification de la classe positive. En effet, seuls 3 modèles sur les 80 n'ont pas constaté d'augmentation de performances avec ces stratégies. Il est donc important de bien déterminer la méthode et les paramètres de ré-échantillonnage avant d'entraîner des modèles. Les deux méthodes les plus couramment utilisées sont le sur-échantillonnage et le sous-échantillonnage aléatoires, méthodes qui ont déjà fait leurs preuves [33].

5.5.2 Étiquetages

À l'aide des figure 5.2 et 5.4, on voit que l'étiquetage qui obtient les meilleures performances est le découpage 1/24, soit des périodes de 7 jours avec les 8 périodes les plus récentes de chaque employé étiquetées comme Période à Risque, les 8 périodes précédentes ignorées puis les 8 périodes précédentes étiquetées comme Période Stable. Ce résultat peut être interprété de la façon suivante : les 8 dernières semaines précédant le départ d'un employé se démarquent bien des 8 avant-avant-dernières semaines. Si un tel modèle obtient de bonnes performances, cela veut dire qu'on est capable de mesurer une chute de satisfaction dans les 8 semaines précédant un éventuel abandon.

Globalement, les modèles qui utilisent les données des 24 dernières semaines de chaque employés (étiquetages $X/24$) ont obtenu les meilleurs résultats. On peut souligner que ces jeux de données ont tendance à être plus grands étant donné que l'horizon temporel est plus élevé. De plus, les deux classes sont séparées par huit semaines contre quatre pour les étiquetages $X/12$ et deux pour les étiquetages $X/6$, ce qui explique en partie pourquoi les deux classes se distinguent mieux.

De surcroît, le nombre de jours dans une période a aussi une influence sur le nombre d'instances. Lorsque la taille d'une période diminue à horizon fixé, le nombre d'instances dans les jeux de données augmente. Ceci expliquerait en partie la baisse des performances entre les découpages 1/24, 2/24 et 4/24. Cette chute de performances se retrouve aussi pour les découpages 1/12, 2/12 et 4/12, mais les découpages 1/6 et 2/6 ne semblent pas toujours affectés par cette tendance.

5.5.3 Résultats à travers les différents clients

D'après la figure 5.5, les performances des modèles varient d'un client à un autre. Tout d'abord, il faut noter que les jeux de données des clients peuvent être assez différents. Ces différences s'expliquent en partie par la diversité du type d'activités, des tailles de structures ou de leur emplacement géographique. Par ailleurs, les usages du logiciel d'où proviennent les données ne sont pas tous identiques : certaines fonctionnalités (pointages, infrastructures...) peuvent ne pas être utilisées. On s'attend donc à ce que le pourcentage de données manquantes pour certaines catégories évoluent en fonction des clients. Cette caractéristiques a un impact direct sur la qualité des données.

Les résultats présentés sur la figure 5.7 permet de visualiser à quel point les variables n'ont pas eu le même impact sur tous les clients. Le signe du coefficient de corrélation n'est pas le même pour tous, ce qui montre que l'on ne peut pas généraliser les résultats mais que ceux-ci doivent être analysés agence par agence. Les écarts-types relevés sur le tableau 5.2 témoignent de la variation de certaines variables à travers les différentes expériences.

Par ailleurs, on peut observer dans le tableau 5.1 à quel point les performances pour chaque client sont éloignées de celle d'un algorithme naïf. La Company_1 est le client qui a obtenu les moins bonnes performances : les résultats du test et d'un algorithme naïf sont proches. Les clients Company_5, Company_9 et Company_10 obtiennent les meilleures performances en regardant la métrique "Signal (%)" définie à la section 3.3. Pour rappel, cette métrique est une échelle linéaire entre un algorithme naïf et un algorithme parfait. Dans ce tableau on remarque que les modèles XGBoost ont des performances presque parfaites (ROCAUC et F_1) à l'entraînement, mais celles-ci chutent beaucoup sur l'ensemble de test : cette observation est caractéristique du sur-apprentissage que l'on peut aussi visualiser sur la figure 5.6.

5.5.4 Différences entre les modèles

À l'aide de la figure 5.5, on remarque que les ordres de grandeurs des résultats des différents modèles sont similaires. Cependant, d'après le tableau 5.1 et la figure 5.5, l'architecture qui a donné les meilleures performances est le modèle XGBoost pour les dix clients. Les modèles non-linéaires ont donné de meilleurs résultats que les modèles linéaires.

La figure 5.6 montre que le sur-apprentissage a été globalement plus important pour les modèles non-linéaires que pour la régression logistique. La moyenne des rapports entre les performances sur l'ensemble d'entraînement et de test pour les méthodes non-linéaires se situe autour de 32% contre 18% pour les régressions logistiques. Cette différence s'explique par la complexité des relations que l'on peut établir entre les variables explicatives et la variable

de sortie avec des relations non-linéaires : la frontière de décision entre les deux classes peut prendre d'autres formes qu'une simple droite.

5.5.5 Variables prédictives

Le tableau 5.2 montre les moyennes des coefficients des régressions logistiques de toutes les compagnies pour tous les découpages essayés. Seules les neuf variables avec la plus grande moyenne en valeur absolue sont conservées afin de visualiser les variables ayant eu le plus d'influence dans la classification. Les coefficients apportent une idée de l'influence de chaque variable explicative sur la variable de sortie. Il faut mentionner que chaque variable est standardisée : les variables ainsi transformées ont donc des valeurs négatives si elles sont inférieures à la moyenne initiale et positives dans l'autre cas. Ainsi les ordres de grandeurs des coefficients associés à ces variables sont comparables entre eux. Si le coefficient d'une variable est positif, cela signifie qu'une valeur plus élevée de celle-ci vient renforcer la probabilité d'appartenance à la classe positive, et inversement.

La variable qui apporte le plus d'information quelle que soit l'architecture utilisée est le nombre de visites (`n_visits`). Avoir un nombre de visites élevé est associé avec une meilleure satisfaction, et un risque de quitter son emploi plus faible. Cette variable est la seule à avoir un résultat qui s'applique à l'ensemble des clients : donner suffisamment de visites à un employé est essentiel afin de conserver les employés.

En regardant toutes les variables du tableau 5.2, on peut montrer que de manière générale, les attributs suivants sont associés à des niveaux de satisfaction plus élevé :

- avoir un nombre de visites élevé (`n_visits`)
- avoir des durées de visites qui changent (`h_std`)
- avoir des visites plus longues (`h_pv`)
- fournir ses indisponibilités de la semaine (`h_unavailability`)

De plus, les employés plus âgés (`age`) semblent avoir un risque plus faible de quitter leur emploi comme l'avait observé [18]. D'après ce même tableau, les zones géographiques des clients visités semblent avoir un impact : le niveau de satisfaction des employés diminue lorsque le pourcentage de personnes mariées (`cli_perc_married`) y est plus élevé ou que le pourcentage de citoyens canadiens diminue (`cli_perc_citizen_can`). Les employés qui habitent dans des zones où le transport en voiture est privilégié (`emp_perc_transp_car`) semblent aussi être plus à risque. Cette dernière variable est statique (constante au cours du temps pour un employé).

D'après les tableaux 5.3 et 5.4, certaines variables ont une importance notable avec les modèles d'arbres mais n'apparaissant pas dans le tableau 5.2 avec les régressions logistiques.

On suppose donc que l'information qui peut être extraite de ces variables n'est pas linéaire avec la variable de sortie. Parmi les variables concernées, on retrouve la distance parcourue par jour (`est_dist_per_day`), le retard moyen par visite (`late_pv`), et la diversité de clients visités (`perc_clients`).

L'objectif de la figure 5.7 est de comparer les impacts des variables entre les différents clients. Certaines variables sont régulières : le nombre de visites, le nombre de visites en périodes de pointe, le nombre d'entrées dans les listes noires, ou l'âge sont des variables dont le signe des coefficients sont réguliers même si l'amplitude varie d'un client à un autre. À l'inverse, certaines variables comme l'influence du retard, et le travail de nuit ou de fin de journée ont montré des résultats très différents selon le client concerné. Cette figure montre l'importance de faire un diagnostic personnalisé pour chaque client, même si certains résultats se généralisent bien. Cette discussion appuie les points soulignés dans la section 5.5.3 sur la différence d'utilisation du logiciel par les différents clients.

Les mesures de distance parcourue par les employés sont une estimation. Elles sont calculées à partir des coordonnées géographiques. La distance réelle dépend de l'agencement des routes dans une zone et du moyen de transport utilisé : voiture, transports en commun, vélo ou encore la marche. Ces données n'étant pas accessibles dans les bases de données étudiées, l'interprétation de la mesure de l'impact des distances parcourues doit être réalisée avec précaution.

Les coefficients de la COMPANY_1 sont presque tous nuls : lors du choix des hyper-paramètres de la régression logistique, les paramètres qui donnaient les meilleures performances sur l'ensemble de validation impliquaient une régularisation *elasticnet* très forte, écrasant alors tous les coefficients proche de zéro. Ceci explique en partie pourquoi les performances obtenues pour ce client sont mauvaises.

5.6 Discussion sur la méthode

5.6.1 Données manipulées

Une première observation concerne la quantité de données. La méthode développée ici requiert la séparation des données en trois jeux distincts : entraînement, validation et test. Beaucoup d'exemples sont nécessaires pour que chaque classe soit bien représentée dans chaque groupe. Dans nos expériences, les plus petits jeux de données comptent 1000 instances, ainsi la sélection des meilleurs hyper-paramètres s'effectue sur quelques 200 exemples, et les performances finales sont estimées sur 300 exemples environ. Les performances de l'ensemble de validation et de test peuvent alors varier fortement, et ne pas représenter fidèlement la réalité du terrain.

5.6.2 Caractéristiques des variables

Certaines variables sont constantes en fonction du temps : le genre de l'employé, les données socio-économiques liées au code postal de l'employé, et le type d'emploi. Au vu des périodes de temps considérées, on considère que l'âge ne varie quasiment pas et fait partie de ces variables statiques. Ces variables ne permettent pas d'expliquer l'aspect dynamique de la satisfaction qui évolue au cours du temps. Cependant, elles peuvent tout de même apporter de l'information : les modèles non-linéaires peuvent croiser ces variables avec d'autres. On pourrait par exemple avoir différentes interprétations pour une période contenant 10 visites selon le genre de l'employé ou son type d'emploi. Ainsi, elles sont conservées dans les modèles.

Par ailleurs, les variables `is_medical` et `is_male` ne peuvent prendre que deux valeurs. Ainsi une fois qu'une variable binaire est utilisée pour faire une séparation sur un noeud d'un arbre, elle ne sera plus utilisée dans les deux sous-arbres créés. La fréquence d'utilisation de ces variables est donc potentiellement plus faible que celle des variables continues ou discrètes qui peuvent prendre de nombreuses valeurs distinctes. Pour cette raison, le coefficient d'importance associé aux variables binaires est donc proche de zéro dans les modèles d'arbres. De plus ces variables ne sont pas standardisées étant données qu'elles sont binaires : l'ordre de grandeur des coefficients qui y sont associés peut donc être légèrement différent des autres variables.

5.6.3 Modèles utilisés

L'explicabilité des variables est plus compliquée à observer avec les modèles d'arbres qu'avec la régression logistique. Les coefficients d'importances des tableaux 5.3 et 5.4 montrent les variables qui ont permis d'apporter le plus d'information dans la création des arbres. Cependant, ces coefficients ne permettent pas de mesurer les impacts de l'évolution d'une variable explicative sur la variable de sortie comme on peut le faire avec les coefficients des régressions logistiques.

Il est possible qu'un entraînement plus approfondi pour les modèles MLP, avec des espaces de paramètres plus grands et un meilleur contrôle du nombre d'itérations puisse améliorer les performances obtenues pour ce type de modèle. Cependant, entraîner ce type d'architecture demande beaucoup de ressources et de temps de calcul. L'utilisation de cartes graphiques permettrait de diminuer le temps de calcul, et certaines plateformes en ligne en proposent gratuitement, mais les données étant confidentielles il n'est pas envisageable de les exporter sur internet. Par ailleurs, il est important d'avoir beaucoup de données pour entraîner ces modèles.

Les paramètres à tester pour un modèle se basent tous sur la stratégie optimale d'équilibrage de classes déterminée sur une architecture avec des paramètres lambda. Dans l'idéal, pour chaque jeu de paramètres, il faudrait re-déterminer la stratégie d'équilibrage optimale, mais cette option n'est pas réalisable étant donné les ressources et le temps disponibles pour ce projet.

Comme le montre la figure 5.2, la métrique F_1 de la classe positive dépend de l'équilibre des classes. Par exemple, le client COMPANY_1 semble obtenir de bons scores avec la F1, mais un mauvais ROCAUC. Il est donc crucial de regarder plusieurs métriques afin de discuter des performances d'une compagnie et de choisir le meilleur modèle.

5.7 Score de satisfaction à l'échelle d'une entreprise

Pour les agences de grande taille, avoir un suivi individuel avec chaque employé est compliqué, d'autant plus que le nombre de gestionnaires peut être élevé. Dans la figure 4.6 de la section 4.2.2, on a montré que les dernières périodes d'un employé actif ne sont pas utilisées pour construire les modèles étant donné que l'étiquette de ces périodes est inconnue. Cependant, une fois les modèles entraînés, il est possible d'assigner une probabilité d'appartenance à la classe positive à ces périodes. Les employés détectés comme étant à risque pourraient être mis en avant auprès des gestionnaires afin que ceux-ci puissent corriger la situation en communiquant avec ces personnes afin de les conserver dans leur structure.

Plusieurs scores de satisfaction peuvent être envisagés :

- considérer le nombre de périodes classifiées comme à risque
- moyenne des probabilités des périodes enregistrées
- ajout de pondérations pour les périodes plus récentes

Un score pourrait alors être établi pour chaque employé encore actif. Cependant, une réflexion doit être menée sur la manière d'utiliser ces informations afin de déterminer un cadre éthique pour l'exploitation de ces données étant donné que ces prédictions sont réalisées au niveau individuel. Une solution est d'agréger tous les scores obtenus sur l'ensemble des employés et de communiquer une moyenne au client : ainsi, la notion individuelle des employés n'est plus présente. Les facteurs ayant une influence sur leurs employés peuvent aussi être communiqués afin qu'ils puissent adapter leur politique d'assignation de visites. On espérerait alors qu'une telle solution améliorerait à court et moyen terme la satisfaction des employés.

CHAPITRE 6 ALGORITHME D'OPTIMISATION

La popularité des soins à domicile au Canada a développé le besoin d'algorithmes d'optimisation de visites pour cette industrie. La société AlayaCare dispose d'un outil d'optimisation répondant au problème de *Home Health Care Routing and Scheduling Problem* (HHCRSP) avec fenêtres de temps. L'objectif est de déterminer les routes que doivent suivre les employés de soins sur un horizon fixé pour voir un ensemble de patients en minimisant un objectif, qui peut prendre en compte le temps de trajet ou la qualité des emplois du temps créés par exemple. La prise en compte des fenêtres de temps impose des créneaux sur lesquels sont disponibles les employés et les patients. Proposé par [39], les solutions sont construites à partir d'une méthode basée sur le partitionnement d'ensembles suivi d'une recherche large de voisinage. Les expériences ont montré que les résultats obtenus permettaient de diminuer le temps de trajet de 37% et d'améliorer significativement la continuité de soins.

À travers cette étude, on a montré que certaines variables liées à la planification de visites ont une influence sur la satisfaction des employés. Chacune d'entre elles ont été identifiées comme ayant un impact positif, négatif ou quasi-nul sur la probabilité de départ d'un employé, avec des différentes amplitudes. Ces résultats peuvent être établis pour chaque client. Prendre en compte ces variables dans la création des routes pourrait améliorer la qualité des emplois du temps créés. On espérerait alors voir la satisfaction globale des employés augmenter avec l'utilisation régulière d'un tel outil. Cette section présente une méthode pour incorporer ces variables quel que soit le client comme un coût additionnel (pénalité ou gain) pour chaque route. La nouvelle formulation des routes n'a pas pu être testée dans le cadre de ce projet, elle reste purement théorique. Dans des travaux futurs, on pourrait s'intéresser à l'analyse des nouvelles solutions, s'assurer qu'il y a un réel apport de qualité aux solutions proposées et que celles-ci sont validées par les structures de soins.

6.1 L'algorithme d'optimisation existant

Le modèle implémenté à AlayaCare est détaillé dans l'article [39]. Il convient de ré-utiliser certaines notations afin d'introduire un nouveau terme dans le coût associé à une route.

6.1.1 Les routes des employés

En reprenant la notation de [39], on définit l'ensemble des patients P et l'ensemble des employés de soins C. L'objectif est de déterminer les routes du personnel infirmier sur un

horizon de H jours (7 dans l'article). Chaque patient $p \in P$ nécessite un nombre de visites n_p sur l'horizon fixé.

On définit l'ensemble des routes Ω. Une route $\omega \in \Omega$ est un ensemble de visites effectuées par un employé $c \in C$ pendant un jour $d \in [1, \ldots, H]$. À chaque route est associé un temps de trajet tt_ω en minutes et un nombre d'heures travaillées len_ω. Les fenêtres de temps étant définies comme des contraintes dures, toutes les routes respectent les disponibilités des employés et des patients. De plus, les routes n'autorisent pas les associations interdites présentes dans les listes noires, et les compétences nécessaires aux patients sont en accord avec les compétences des employés.

Certaines préférences optionnelles liées aux employés peuvent être communiquées par le patient : sexe, langues parlées, ou si celui-ci est non-fumeur par exemple. Le nombre de préférences optionnelles non-satisfaites pour une route ω est définit par $\bar{o}_\omega$.

La variable binaire $a_{\omega,p}$ vaut 1 si le patient p est visité par la route ω, sachant qu'un patient ne peut pas être vu plus d'une fois par jour.

En connaissant les visites qui ont eu lieu dans la semaine précédant l'horizon considéré, on peut dénombrer le nombre de fois qu'un patient p a été vu par l'employé de soins c. Cette valeur est stockée dans le paramètre $\text{CC}_{p,c}$ et permet d'avoir une mesure de la continuité de soins.

6.1.2 Critères de coût pour une route

Le coût d'une route c_ω est défini de la manière suivante par [39] :

$$c_\omega = \gamma_1 \cdot \bar{o}_\omega + \gamma_2 \cdot tt_\omega + \gamma_3 \cdot \sum_{p \in P} a_{\omega,p} \cdot f_1\left(\text{CC}_{p,c}\right) + \gamma_4 \cdot f_2\left(\text{len}_\omega\right)$$

Les poids vérifient $\forall i \in \{1,2,3,4\}, \gamma_i \geq 0$. Le premier terme prend en compte les préférences optionnelles non-satisfaites de chaque route. Le deuxième est relatif au temps de trajet. Le troisième permet d'avoir une mesure de la continuité de soins avec :

$$f_1\left(\text{CC}_{p,c}\right) = \begin{cases} 1 & \text{si } \text{CC}_{p,c} = 0 \\ \frac{2}{3} & \text{si } 1 \leq \text{CC}_{p,c} \leq 2 \\ \frac{1}{3} & \text{sinon} \end{cases}$$

Enfin, le quatrième terme ajoute une pénalité en cas de non respect du minimum/maximum horaire quotidien. En définissant $\underline{w}_d$ et $\overline{w}_d$ respectivement comme le minimum et le maximum

horaire pour le jour $d \in [1, \ldots, H]$, [39] propose :

$$f_2\left(\text{len}_\omega\right) = \begin{cases} \underline{w}_d - \text{len}_\omega & \text{si len}_\omega < \underline{w}_d \\ \text{len}_\omega - \overline{w}_d & \text{si len}_\omega > \underline{w}_d \\ 0 & \text{sinon} \end{cases}$$

6.1.3 Fonction objectif

Les variables de décision utilisées sont :

- x_ω : binaire qui vaut 1 si ω est active, 0 sinon
- o_c : minutes de dépassement de la capacité horaire hebdomadaire maximale de l'employé c
- u_c : minutes d'inactivité de l'employé c
- z_p : nombre de visites du patient p non affectées

La fonction objectif est alors la suivante :

$$\min \sum_{\omega \in \Omega} c_\omega x_\omega + \beta_1 \cdot \sum_{c \in C} (o_c + u_c) + \beta_2 \cdot \sum_{p \in P} z_p$$

Le premier terme est lié au coût d'une route précédemment défini, le deuxième terme permet de mesurer les dépassements horaires hebdomadaires et le temps d'inactivité des employés, et le troisième est lié au nombre de visites non affectées. Les poids $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \beta_1, \beta_2)$ sont tous positifs, et ont été choisis en collaboration avec AlayaCare selon les préférences d'un client.

6.1.4 Méthode de résolution

La résolution du problème commence par la création d'une solution initiale à l'aide d'une heuristique simple. Il suffit d'insérer les visites les plus longues aux positions de coût minimal. Pendant un certain nombre d'itération N, des opérateurs de destruction et de réparation de visites sont appliqués sur la solution courante. Ceux-ci sont présentés au paragraphe suivant. À chaque fois, si la nouvelle solution est meilleure que la précédente, elle est conservée. Sinon, elle est conservée sous une certaine probabilité. Ce cycle est répété plusieurs fois, mais en repartant d'une solution établie en relaxant la contrainte d'intégrité de x_ω. Une procédure est ensuite appliquée pour rétablir une solution entière. Un critère d'arrêt en temps ou en nombre d'itérations interrompt l'algorithme et retourne la meilleure solution trouvée pendant ce processus.

Les différents opérateurs de destruction et de réparation utilisés sont définis par Ropke [40]

et Shaw [41] :

- *Random Removal* : retire une visite aléatoirement
- *Worst Removal* : retire les visites qui détériorent le plus la fonction objectif
- *Related Removal* : retire les visites à partir d'une mesure de *relatedness* calculée sur les routes et les coûts d'insertion des visites aux employés
- *Basic Greedy Heuristic* : insère la visite qui détériore le moins la fonction objectif
- *Regret-2* et *Regret-3* : ajoutent des visites en anticipant les insertions futures

Deux opérateurs de destruction et un de reconstruction supplémentaires sont définis spécifiquement pour ce problème [39] :

- *Service Removal* : choisit aléatoirement des patients et retire toutes ses visites programmées
- *Flexible Avail Removal* : retire les visites des patients en commençant avec ceux qui ont la plus grande flexibilité
- *Random Service* : assignation de coût minimal de visites choisies aléatoirement

Retirer toutes les visites d'un patient d'un coup peut avoir un impact positif sur une solution. Celles-ci pourraient être réaffectées à un employé qui garantirait une meilleure continuité de soin par exemple. Enlever les visites des patients les plus flexibles donne plus d'options lors de la reconstruction. Diversifier les opérateurs de destruction et de reconstruction permet de varier les solutions créées à chaque itération afin de ne pas retomber systématiquement sur les mêmes. L'algorithme utilise donc 5 opérateurs de destruction et 4 de reconstruction.

6.2 Notre apport

Nous proposons une nouvelle formule pour le coût c_ω d'une route qui prend en compte les variables identifiées comme ayant un impact sur la satisfaction des employés. Pour cela, il faut tout d'abord adapter les variables qui étaient relevées sur des périodes à des routes.

6.2.1 Variables d'intérêt

Parmi les variables utilisées dans notre étude présentées dans les tableaux A.1 et A.2, seules les suivantes ont été identifiées comme pouvant être inclues dans le modèle :

- `h_pv`
- `h_std`
- `perc_servcodes`
- `weekend_perc`
- `night_perc`
- `evening_perc`

— `peak_percentage`

Les expériences ont montré que le nombre de visites est fortement corrélé au nombres d'heures len_ω. Le deuxième facteur étant déjà pris en compte dans la formule de coût d'une route, il ne sera pas nécessaire d'ajouter un critère pour le nombre de visites. Les variables liées aux listes noires, aux disponibilités des employés et des distances parcourues ont déjà été implémentées dans l'algorithme de création de routes comme une contrainte dure, on est donc assuré qu'elles seront respectées. Les variables concernant le retard aux visites, aux visites annulées, ou les caractéristiques des employés n'ont pas d'applicabilité dans cette méthode. Il a été choisi de ne pas inclure les variables socio-démographiques liées aux patients car elles pourraient entraîner de la discrimination dans le choix des visites affectées.

6.2.2 Application à une route

Variables définies pour une route

Les variables précédentes s'appliquaient sur une période de plusieurs jours dans notre étude. La modélisation de ce problème d'optimisation prend en compte des routes journalières. Ainsi, on définit les paramètres suivants afin d'adapter les variables d'intérêt de la section précédente à une route $\omega \in \Omega$:

— $\lambda_{\omega,1}$: durée moyenne des visites de la route
— $\lambda_{\omega,2}$: écart type des durées des visites de la route
— $\lambda_{\omega,3}$: mesure de la diversité des services dans la route (définie dans les prochains paragraphes)
— $\lambda_{\omega,4}$: vaut 1 si la route a lieu pendant un jour de fin de semaine, 0 sinon
— $\lambda_{\omega,5}$: nombre de visites qui ont lieu pendant la nuit
— $\lambda_{\omega,6}$: nombre de visites qui ont lieu pendant la fin de journée
— $\lambda_{\omega,7}$: nombre de visites qui ont lieu pendant les heures de pointe

Les horaires correspondant aux périodes de la journée et aux horaires de pointes ont été présentées aux figures 4.2 et 4.3 respectivement. Pour évaluer la diversité de services donnés pendant une route, il convient de définir l'ensemble S qui contient l'ensemble des services possibles. Pour chaque route $\omega \in \Omega$, on calcule n_ω^s le nombre de services distincts associés aux visites, et $n_\omega^{visites}$ le nombre de visites dans ω. On définit ainsi :

$$\lambda_{\omega,3} = \frac{n_\omega^s}{\min(n_\omega^{visites}, |S|)}$$

La mesure de diversité $\lambda_{\omega,3}$ est majorée par 1 puisque toutes les visites sont associées à un service. Elle est maximale lorsque :

— chaque visite a un service distinct des autres lorsque $n_\omega^{visites} < |S|$
— tous les services sont représentés au moins une fois lorsque $n_\omega^{visites} \geq |S|$

Nouvelle formulation de la fonction de coût

L'objectif de cette méthode est d'ajouter les variables précédemment définies comme un coût supplémentaire au coût c_ω défini par [39]. Une première idée est de prendre une somme pondérée de chaque variable dont les poids seraient fixés manuellement, le facteur supplémentaire serait alors de la forme $\sum_{i=1}^{7} \gamma_i \cdot g_i(\lambda_{\omega,i})$. En plus des 6 poids de la méthode originale à déterminer $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \beta_1, \beta_2)$, cette solution implique de devoir fixer les 7 coefficients γ_i en plus, soit 13 paramètres au total. Ce nombre pourrait être réduit en regroupant certaines variables, comme les variables $\{\lambda_1, \lambda_2\}$ qui se rapportent à la durée des visites ou les variables $\{\lambda_5, \lambda_6, \lambda_7\}$ qui concernent les plages horaires des visites, on aurait alors plus que 4 coefficients γ à fixer. La méthode retenue nécessite qu'un paramètre γ à fixer manuellement. D'autres poids α_i devront être déterminés pour régler l'importance relative des variables entre elles, mais ceux-ci seront choisis par l'expérience et la modélisation du problème de satisfaction abordé dans ce projet. Nous proposons la nouvelle formulation suivante pour le coût associé à une route ω, noté c'_ω :

$$c'_\omega = c_\omega + \gamma_5 \cdot \sum_{i=1}^{7} \alpha_i \cdot g_i(\lambda_{\omega,i})$$

Le poids γ_5 représente l'importance du nouveau terme par rapport aux quatre autres définis dans la fonction de coût c_ω de base, et est positif. Pour tout $i \in [1, \ldots, 7]$, α_i représente le poids relatif de chaque variable, et la fonction g_i correspond à la transformation associée à la variable $\lambda_{\omega,i}$.

6.2.3 Détermination des fonctions g_i

Les 7 variables sont toutes numériques. On propose deux transformations possibles : les fonctions u et v qui correspondent à la normalisation min-max et à la standardisation respectivement.

Normalisation min-max

Afin que les variables évoluent dans le même intervalle, on définit la fonction u qui transforme linéairement les réalisations $x \in F$ avec $F \subset \mathbf{R}$ d'une variable X entre 0 et 1 :

$$u_F(x) = \begin{cases} \frac{x - \min_F(X)}{\max_F(X) - \min_F(X)} & \text{si } \max_F(X) - \min_F(X) > 0 \\ 0 & \text{sinon} \end{cases}$$

Ainsi, u_F est maximal ($u_F(x) = 1$) lorsque x atteint le maximum de X sur F, et est minimal ($u_F(x) = 0$) lorsqu'il en atteint le minimum. En définissant Ω_c l'ensemble des routes associées à l'employé $c \in C$ avec $\Omega_c \subset \Omega$, la nouvelle fonction de coût c'_ω aura la forme suivante :

$$c'_\omega = c_\omega + \gamma_5 \cdot \sum_{i=1}^{7} \alpha_i \cdot u_{\Omega_c}(\lambda_{\omega,i})$$

Standardisation

Une autre solution est de standardiser la variable X en une nouvelle variable X' : en lui recoupant sa valeur moyenne et en le divisant par son écart-type, on a alors $E(X') = 0$ et $Var(X') = 0$. On définit alors la fonction v qui pour une réalisation $x \in F$ avec $F \subset \mathbf{R}$ d'une variable X associe sa valeur standardisée :

$$v_F(x) = \begin{cases} \frac{x - E_F(X)}{\sqrt{\mathrm{Var}_F(X)}} & \text{si } \mathrm{Var}_F(X) > 0 \\ 0 & \text{sinon} \end{cases}$$

En reprenant les ensembles Ω_c pour $c \in C$, la nouvelle fonction de coût s'écrira alors :

$$c'_\omega = c_\omega + \gamma_5 \cdot \sum_{i=1}^{7} \alpha_i \cdot v_{\Omega_c}(\lambda_{\omega,i})$$

6.2.4 Détermination des poids α_i

Dans le chapitre 4, nous avons proposé une manière de modéliser le départ des employés pour un problème de classification binaire à deux classes. On a ainsi déterminé une variable de sortie y qui vaut 1 lorsqu'une période est étiquetée comme appartenant à la classe Période à Risque et 0 sinon. L'idée est de se servir de cette modélisation pour déterminer les poids α_i liés à chaque variable. Ces poids seront déterminés à partir des données historiques d'un client, soit toutes les visites qui ont eu lieu en amont de l'horizon étudié.

Création d'un jeu de données adapté

Les variables utilisées dans cette section sont définies pour une route. Afin que les résultats de la méthode développée dans les chapitres précédents soient utilisables, il faut paramétrer la taille d'une période à un jour. Ainsi chaque exemple de la base de données correspondra à une route pour un employé pendant un certain jour. Les routes proches du départ d'un employé seront étiquetées comme appartenant à la classe positive tandis que les routes appartenant au passé plus lointain appartiendront à la classe négative. Le nombre de périodes optimal de

périodes qu'il faut sélectionner pour chaque classe peut être déterminé par l'expérience avec une méthode semblable à celle proposée au chapitre 5. On avait ainsi réussi à déterminer les paramètres de découpage qui assuraient les meilleures performances. Pour chaque route considérée, il faudra relever les variables $\bar{\lambda}_{\omega,i}$, $i \in [1, \ldots, 7]$ présentées à la section 6.2.2. La notation avec une barre permet de distinguer les variables historiques de celles de l'horizon d'étude.

Coefficient de corrélation de Pearson

Une fois les variables relevées, on est capable de calculer le coefficient de corrélation linéaire de Pearson présenté à la section 3.1.2. Ainsi pour tout $i \in [1, \ldots, 7]$, le coefficient $r_{\bar{\lambda}_i,y}$ donne deux informations : son signe indique si la variable doit être considérée comme une pénalité ou un gain, et son amplitude montre l'importance de la corrélation linéaire. Ainsi, une première solution est de considérer :

$$\alpha_i^{(1)} = \frac{r_{\bar{\lambda}_i,y}}{\sum_{j=1}^{7} |r_{\bar{\lambda}_j,y}|}$$

La normalisation par la somme des coefficients permet de garantir $\sum_{i=1}^{7} |\alpha_i^{(1)}| = 1$ pour que le facteur que l'on cherche à ajouter à la fonction de coût prenne des valeurs cohérentes à travers les clients. Pour rappel, les coefficients α_i sont déterminés à partir des données historiques, donc antérieures à l'horizon étudié.

Certaines variables peuvent être faiblement corrélées à la variable de sortie. Une manière d'écraser les coefficients liés à ces variables est d'utiliser le carré du coefficient de corrélation. Il faut cependant penser à conserver le signe de $r_{\bar{\lambda}_i,y}$ puisque celui-ci indique si la variable correspond à un gain ou une pénalité dans le coût de la route. On a alors :

$$\alpha_i^{(2)} = \text{signe}(r_{\bar{\lambda}_i,y}) \cdot \frac{r^2_{\bar{\lambda}_i,y}}{\sum_{j=1}^{7} r^2_{\bar{\lambda}_j,y}}$$

La somme des valeurs absolues de ces nouveaux coefficients est encore égale à 1. En utilisant la fonction u définie précédemment sur l'ensemble Ω_c, on vérifie l'inégalité suivante pour les deux dernières définitions des poids α_i :

$$\begin{aligned} |\sum_{i=1}^{7} \alpha_i \cdot u_{\Omega_c}(\lambda_{i,\omega})| &\leq \sum_{i=1}^{7} |\alpha_i \cdot u_{\Omega_c}(\lambda_{i,\omega})| \\ &\leq \sum_{i=1}^{7} |\alpha_i| \\ &\leq 1 \end{aligned}$$

Ainsi, ce nouveau terme est compris entre -1 et 1. Une valeur de 1 indique que la route a tendance à favoriser le départ d'un employé, le coût de la route augmente. Une valeur de

-1 indique que la route est considérée comme ayant un impact positif sur la satisfaction de l'employé, le coût de la route diminue. La fonction de coût proposée aura alors un des deux formes suivantes selon la définition des α_i choisie :

$$c_\omega^{(1)} = c_\omega + \gamma_5 \cdot \sum_{i=1}^{7} \frac{r_{\bar{\lambda}_i,y}}{\sum_{j=1}^{7} |r_{\bar{\lambda}_j,y}|} \cdot u_{\Omega_c}(\lambda_{\omega,i})$$

$$c_\omega^{(2)} = c_\omega + \gamma_5 \cdot \sum_{i=1}^{7} \text{signe}(r_{\bar{\lambda}_i,y}) \cdot \frac{r_{\bar{\lambda}_i,y}^2}{\sum_{j=1}^{7} r_{\bar{\lambda}_j,y}^2} \cdot u_{\Omega_c}(\lambda_{\omega,i})$$

Coefficients d'une régression logistique

Une autre méthode envisageable est de récupérer directement les coefficients d'une régression logistique. Une fois de plus, le signe des coefficients et leur amplitude donne une idée de l'influence de chacune des variables. L'ajout de régularisation (L_1, L_2 ou *elasticnet*) permet de bénéficier de l'effet d'écrasement des poids faibles mentionnés plus haut.

Ainsi, l'idée est de lancer une régression logistique avec les 7 variables d'intérêt $\bar{\lambda}_1, \ldots, \bar{\lambda}_7$ (avec ou sans régularisation) sur le jeu de données décrit plus haut. Il est important de standardiser les variables dans la régression logistique afin que les coefficients soient comparables entre eux. Les coefficients $\bar{\beta}_1^{LR}, \ldots, \bar{\beta}_7^{LR}$ déterminés pour chaque variable sont relevés et appliqués directement tel que $\forall i \in [1, \ldots, 7]$:

$$\alpha_i = \bar{\beta}_i^{LR}$$

Le terme de coût s'inspire du modèle logit. On utilise les valeurs standardisées des variables $\lambda_1, \ldots, \lambda_7$ à l'aide de v_{Ω_c}, que l'on pondère par les coefficients $\bar{\beta}_i^{LR}$. En prenant ensuite l'image de cette somme pondérée par la fonction sigmoïde, la quantité ainsi créée se situe entre 0 et 1 et est d'autant plus grande que la probabilité d'appartenance à la classe positive est forte. Pour que cette quantité soit comprise entre -1 et 1 comme à la partie précédente, on peut considérer la transformation suivante :

$$\begin{array}{rccccc} & 0 & \leq & \text{sigm}\left(\sum_{i=1}^{7} \bar{\beta}_i^{LR} \cdot v_{\Omega_c}(\lambda_{\omega,i})\right) & \leq & 1 \\ \Longrightarrow & -1 & \leq & 2 \cdot \text{sigm}\left(\sum_{i=1}^{7} \bar{\beta}_i^{LR} \cdot v_{\Omega_c}(\lambda_{\omega,i})\right) - 1 & \leq & 1 \end{array}$$

avec $\forall x \in \mathbf{R}$:

$$\text{sigm}(x) = \frac{1}{1+e^x}$$

Cette nouvelle quantité est proche de -1 lorsque la probabilité d'appartenance à la classe

négative est certaine, le coût de la route sera alors plus faible. De manière analogue, le coût de la route augmente lorsque la probabilité d'appartenance à la classe positive augmente, jusqu'au maximum de 1. Une valeur de 0 correspond à une incertitude parfaite entre les deux classes, n'ayant aucun impact sur la fonction de coût. La formule de coût proposée est donc :

$$c_\omega^{(3)} = c_\omega + \gamma_5 \cdot \left[2 \cdot \text{sigm} \left(\sum_{i=1}^{7} \bar{\beta}_i^{LR} \cdot v_{\Omega_c}(\lambda_{\omega,i}) \right) - 1 \right]$$

6.2.5 Détermination du poids γ_5

Le coefficient γ_5 pondère l'apport du nouveau terme de coût par rapport aux autres termes. Afin de simplifier le réglage manuel de γ_5, la pénalité a été définie de telle sorte qu'elle évolue dans l'intervalle $[-1, 1]$ quelle que soit la fonction de coût choisie $c_\omega^{(1)}$, $c_\omega^{(2)}$ ou $c_\omega^{(3)}$. L'article [39] suggère de fixer les poids manuellement à travers les expériences et les discussions avec le client concerné. Cependant, [42] propose une approche hiérarchique des critères de la fonction multi-objectif. Une liste de critères est fournie au client qui doit ordonner l'importance de chacun. L'avantage de cette méthode est qu'elle devient très intuitive pour le client : la recherche manuelle des poids optimaux peut être très longue et fastidieuse, tandis qu'une simple hiérarchisation des critères est facile à établir et est aisément applicable d'un client à l'autre. Ainsi l'ajout du critère « Critères liés au risque de départ » dans cette liste laisse la liberté au client de choisir l'importance de celui-ci.

6.3 Discussion de la méthode

L'avantage principal de cette méthode est qu'elle qu'il n'y a qu'un paramètre supplémentaire à fixer : γ_5. Il faut mentionner qu'une recherche de découpage optimal est nécessaire afin de pouvoir déterminer les coefficients de corrélation ou les coefficients des régressions logistiques. Cette méthode vient aussi avec l'inconvénient suivant : il n'est pas possible de régler l'influence des nouvelles variables séparément, les paramètres α sont déterminés par l'expérience seulement.

Dans les deux méthodes (Pearson et régression logistique), les valeurs des α_i sont déterminées à partir des données historiques sur l'ensemble des employés. Or dans la formulation de coût, la normalisation min-max et la standardisation s'appuient sur l'ensemble des routes de chaque employé séparément Ω_c sur l'horizon H. Ce choix fait que les coûts ne vont pas être distribués sur les meilleures routes parmi tous les employés mais parmi chaque employé individuellement.

Dans un ensemble Ω_c, il est possible que la variance d'une des variables soit nulle. Les défini-

tions des fonctions u et v permettent d'annuler l'importance de celles-ci. Dans le cas extrême où toutes les variables sont de variance nulle, alors tous les coefficients de corrélations sont nuls et la définition des $\alpha_i^{(1)}$ et $\alpha_i^{(2)}$ ne sont plus applicables. Il faut alors vérifier préalablement qu'au moins une variable est de variance non-nulle. Dans le cas contraire, il faudra forcer le nouveau terme de coût à zéro. Ce cas peut arriver par exemple si un employé n'a qu'une seule route, soit $\exists c \in C$ tel que $\text{card}(\Omega_c) = 1$.

Dans le cas où toutes les corrélations sont très faibles, la méthode des coefficients de la régression logistique permet de faire ressortir l'absence de signal entre les variables : la probabilité d'appartenance à la classe positive vaudra 0.5, et donc le facteur additionnel dans la fonction de coût sera nul. Par contre, la méthode des coefficients de Pearson n'est pas sensible au cas où les corrélations sont faibles : la normalisation des coefficients $\alpha_i^{(1)}$ et $\alpha_i^{(2)}$ fait que le coût additionnel pourra quand même prendre des valeurs entre -1 et 1.

La méthode qui utilise les coefficients de corrélation linéaire de Pearson n'est pas affectée par la présence de multicolinéarité entre les variables, seule la corrélation avec la variable de sortie est mesurée. Ainsi pour la méthode avec la régression logistique, il est possible que les signes ou les amplitudes des coefficients ne reflètent pas le véritable apport d'une variable.

CHAPITRE 7 CONCLUSION

Ce projet propose une étude quantitative des causes du départ des employés dans le domaine des soins à domicile à partir de données industrielles fournies par AlayaCare, et une application dans un algorithme d'optimisation. Ce problème a été abordé comme un problème de classification à deux classes. La modélisation essayée consiste à différencier les périodes proches du départ d'un employé des périodes plus anciennes. Avec une telle comparaison, on est capable d'expliquer les facteurs qui ont évolué au cours du temps. Différents découpages ont été essayés en faisant varier l'horizon de données afin de valider la période optimale du problème. Une bonne compréhension des données a été nécessaire afin de pouvoir les utiliser pour entraîner des modèles. En déterminant la meilleure stratégie d'échantillonnage et le meilleur type de modèle pour chaque compagnie, on est capable d'analyser les performances d'une solution proposée tout en identifiant les variables explicatives importantes. Le seul résultat qui s'est généralisé sur l'ensemble des agences de soins à domicile est qu'un faible nombre de visites est caractéristique de l'approche du départ d'un employé. Les résultats des différents modèles ont été croisés en utilisant plusieurs métriques afin de renforcer les conclusions émises. Enfin, une nouvelle formulation de la fonction de coût d'une visite de l'algorithme d'optimisation de visites d'AlayaCare a été proposée. Celle-ci prend en compte les variables importantes identifiées dans les expériences réalisées au cours de cette étude. Différentes formulations ont été explorées ainsi que les manières de paramétrer le nouveau terme.

Les données utilisées ont pour la plupart été entrées manuellement, elles sont donc sensibles aux erreurs humaines. De plus, chaque client a une utilisation différente du logiciel d'AlayaCare. La qualité des données ou même la présence de certaines variables ne sont pas garanties. Par ailleurs, il faut une certaine quantité de données historiques pour avoir un volume de données d'entraînement suffisant, il n'est donc pas possible d'appliquer cette méthode pour une agence créée il y a quelques mois.

Les agences de soins à domicile peuvent être très différentes selon leur pays, leur taille ou leur activité. Les conclusions peuvent donc varier d'une structure à une autre, et les modèles doivent être ré-entraînés à chaque nouvelle structure. De plus, les données socio-économiques sont généralement accessibles à l'échelle d'un pays, les variables sont donc différentes entre les structures.

Plusieurs hypothèses fortes ont été formulées pour cette modélisation. Tout d'abord, l'approche avec un problème de classification est discutable, la satisfaction n'est pas un état

binaire mais continu dont la variation peut être irrégulière. En pratique, il est compliqué d'avoir accès au niveau de satisfaction d'un employé en continu, de manière qualitative ou quantitative. Des simplifications ont donc dû être formulées afin de modéliser ce problème.

Cette limitation a donc un impact sur les performances que l'on peut attendre d'une solution à ce problème. La satisfaction d'une personne ne peut pas être complètement expliquée à travers ses données d'activité. Mesurer toutes les variables qui ont une influence sur le niveau de satisfaction d'une personne en continu n'est pas possible, on ne peut donc expliquer qu'une partie de ce phénomène. Il faut cependant noter que les effets liés aux visites n'est pas négligeable et ne doit pas être oublié : ce sont des données nécessaires mais pas suffisantes.

L'accessibilité à des puissances de calcul plus élevées aurait permis d'augmenter l'espace de recherche des hyper-paramètres des modèles en particulier pour les réseaux de neurones. D'une part, le nombre d'hyper-paramètres à tester est particulièrement grand, notamment la quantité et la taille des couches cachées ainsi que le réglage du taux d'apprentissage. De plus, le nombre d'itérations nécessaires pour atteindre la convergence n'est pas connu à l'avance. Les données étant confidentielles, nous n'avons pas pu avoir accès aux ressources disponibles en ligne.

D'autres modélisations de la satisfaction d'un employé pourraient être explorées dans des travaux futurs. L'approche avec des séries temporelles est envisageable et pourrait donner des résultats dynamiques. D'autres types de modèles pourraient être implémentés, tout en gardant un bon compromis entre performance et explicabilité. Le suivi des employés pourrait être plus diversifié à travers le logiciel. En ajoutant des fonctionnalités de collecte de données, on pourrait récupérer de l'information sur la qualité d'une visite par exemple, ou avoir des réponses à des questionnaires de satisfaction envoyés de manière régulière.

Si un client inconnu apparaît dans la base de données, certaines hypothèses de départ peuvent être faites en fonction des résultats déjà obtenus. On peut par exemple comparer les caractéristiques de ce nouveau client avec ceux sur lesquels la méthode a déjà été appliquée. On pourrait alors s'attendre à avoir des résultats semblables. Avec plus de données, il serait envisageable d'essayer de regrouper les clients ayant des similarités pour voir si les résultats obtenus sont comparables. Ces groupes de clients pourraient être réalisés selon la localité, le type de services offerts ou la taille de l'agence par exemple. Ainsi les résultats seraient regroupés par type de structure, afin de raffiner cette analyse.

Certaines précautions doivent être prises avant d'utiliser la nouvelle fonction de coût proposé dans le chapitre 6. Tout d'abord, il faudrait lancer une batterie de tests sur des vraies données afin de valider l'apport du nouveau terme et s'assurer que les routes sélectionnées correspondent aux critères identifiés préalablement. Les nouvelles solutions produites par

l'algorithme devront ensuite être analysées à travers différentes métriques pour s'assurer qu'elles sont acceptables et même meilleures que les anciennes. Une phase supplémentaire devrait être réalisée auprès des employés pour s'assurer que les facteurs identifiés par nos expériences correspondent à leurs préférences : se fier uniquement aux données ne serait pas prudent.

RÉFÉRENCES

[1] Association canadienne de soins et services à domicile (ACSSD), Association des infirmières et infirmiers du Canada (AIIC), Collège des médecins de famille du Canada (CMFC). (2016) Plan national pour de meilleurs soins à domicile au canada.

[2] B. Bernard. (2019) La pénurie de main-d'œuvre en soins infirmiers persiste au canada. [En ligne]. Disponible : https://www.hiringlab.org/fr-ca/blog/2019/12/04/infirmiers-au-canada/

[3] H. Gilmour, "Besoins insatisfaits en matière de soins à domicile au canada," *Rapports sur la santé / Statistique Canada, Centre canadien d'information sur la santé*, vol. 29, p. 3–13, 2018.

[4] Home Care Pulse. (2019) 2019 home care benchmarking study, spring edition.

[5] C. Dall'Ora *et al.*, "Association of 12 h shifts and nurses' job satisfaction, burnout and intention to leave : findings from a cross-sectional study of 12 european countries," *BMJ Open*, vol. 5, n°. 9, 2015.

[6] J. Shiao *et al.*, "Factors predicting nurses' consideration of leaving job (also to be considered for mini-symposium : Early detection and management of workers under stress)," *Occupational and Environmental Medicine*, vol. 71, n°. Suppl 1, p. A94–A95, 2014.

[7] C.-C. Ma, M. E. Samuels et J. W. Alexander, "Factors that influence nurses' job satisfaction," *JONA : The Journal of Nursing Administration*, vol. 33, n°. 5, p. 293–299, 2003.

[8] N. G. Castle *et al.*, "Job Satisfaction of Nurse Aides in Nursing Homes : Intent to Leave and Turnover," *The Gerontologist*, vol. 47, n°. 2, p. 193–204, mai 2007.

[9] E. Chang *et al.*, "Measuring job satisfaction among healthcare staff in the United States : a confirmatory factor analysis of the Satisfaction of Employees in Health Care (SEHC) survey," *International Journal for Quality in Health Care*, vol. 29, n°. 2, p. 262–268, 2017.

[10] M. D. McHugh *et al.*, "Nurses' widespread job dissatisfaction, burnout, and frustration with health benefits signal problems for patient care," *Health affairs (Project Hope)*, vol. 30, n°. 2, p. 202–210, 2011.

[11] M. Merces *et al.*, "Prevalence and factors associated with burnout syndrome among primary health care nursing professionals : A cross-sectional study," *International Journal of Environmental Research and Public Health*, vol. 17, p. 474, 2020.

[12] E. C. Nelson *et al.*, "The relationship between diagnosed burnout and sleep measured by activity trackers : Four longitudinal case studies," dans *Body Area Networks : Smart IoT and Big Data for Intelligent Health Management*, L. Mucchi *et al.*, édit. Cham : Springer International Publishing, 2019, p. 315–331.

[13] W. V. P., W. Y. D. et O. Jodi, "The work-life interface : a critical factor between work stressors and job satisfaction," vol. 48, n°. 4, p. 880–897, 2019.

[14] C. Maslach, S. Jackson et M. Leiter, *The Maslach Burnout Inventory Manual*, 1997, vol. 3, p. 191–218.

[15] R. Alpern *et al.*, "Development of a brief instrument for assessing healthcare employee satisfaction in a low-income setting," *PloS one*, vol. 8, n°. 11, Nov 2013.

[16] C. W. Mueller et J. C. McCloskey, "Nurses' job satisfaction : a proposed measure." *Nursing research*, 1990.

[17] M. Kramer et L. P. Hafner, "Shared values : Impact on staff nurse job satisfaction and perceived productivity." *Nursing research*, 1989.

[18] J. Adriaenssens, V. D. Gucht et S. Maes, "Determinants and prevalence of burnout in emergency nurses : A systematic review of 25 years of research," *International Journal of Nursing Studies*, vol. 52, n°. 2, p. 649 – 661, 2015.

[19] L. Mikyoung et J. Keum-Seong, "Nurses' emotions, emotional labor, and job satisfaction," vol. 13, n°. 1, p. 16–31, Jan 2019.

[20] A. E. Tourangeau *et al.*, "Measurement of nurse job satisfaction using the mccloskey/mueller satisfaction scale," *Nursing Research*, vol. 55, n°. 2, p. 128–136, 2006.

[21] E. T. Lake, "Development of the practice environment scale of the nursing work index," *Research in Nursing & Health*, vol. 25, n°. 3, p. 176–188, 2002.

[22] S.-H. Bae et D. Fabry, "Assessing the relationships between nurse work hours/overtime and nurse and patient outcomes : Systematic literature review," *Nursing Outlook*, vol. 62, n°. 2, p. 138–156, Mar 2014.

[23] B. Karsh, B. Booske et F. Sainfort, "Job and organizational determinants of nursing home employee commitment, job satisfaction and intent to turnover," *Ergonomics*, vol. 48, n°. 10, p. 1260–1281, 2005.

[24] M. A. Al Maqbali, "Factors that influence nurses' job satisfaction : a literature review," *Nursing management*, vol. 22, n°. 2, 2015.

[25] C.-Y. J. Peng, K. L. Lee et G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, n°. 1, p. 3–14, 2002.

[26] R. Manongdo et G. Xu, “Applying client churn prediction modeling on home-based care services industry,” dans *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, Nov 2016, p. 1–6.

[27] A. Y. Ng, “Feature selection, l1 vs. l2 regularization, and rotational invariance,” dans *Proceedings of the Twenty-First International Conference on Machine Learning*, ser. ICML ’04. New York, NY, USA : Association for Computing Machinery, 2004, p. 78.

[28] C. Strobl, J. Malley et G. Tutz, “An introduction to recursive partitioning : rationale, application, and characteristics of classification and regression trees, bagging, and random forests,” *Psychological methods*, vol. 14, n°. 4, p. 323–348, Dec 2009.

[29] J. Brownlee. (2016) A gentle introduction to xgboost for applied machine learning. [En ligne]. Disponible : https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[30] B. Gregory, “Predicting customer churn : Extreme gradient boosting with temporal data,” *arXiv preprint arXiv :1802.03396*, 2018.

[31] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[32] D. Golmohammadi et N. Radnia, “Prediction modeling and pattern recognition for patient readmission,” *International Journal of Production Economics*, vol. 171, p. 151 – 161, 2016.

[33] G. E. Batista, R. C. Prati et M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *SIGKDD Explor. Newsl.*, vol. 6, n°. 1, p. 20–29, juin 2004.

[34] N. V. Chawla *et al.*, “Smote : synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, p. 321–357, 2002.

[35] P. Chapman *et al.*, “Crisp-dm 1.0 : Step-by-step data mining guide,” 2000.

[36] I. Tomek, “Two modifications of cnn,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, n°. 11, p. 769–772, Nov 1976.

[37] T. Shah. (2017) About train, validation and test sets in machine learning. [En ligne]. Disponible : https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

[38] Statistics Canada. (2016) Census profile, 2016 census.

[39] F. Grenouilleau *et al.*, “A set partitioning heuristic for the home health care routing and scheduling problem,” *European Journal of Operational Research*, vol. 275, n°. 1, p. 295–303, 2019.

[40] S. Ropke et D. Pisinger, “An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows,” *Transportation science*, vol. 40, n°. 4, p. 455–472, 2006.

[41] P. Shaw, “Using constraint programming and local search methods to solve vehicle routing problems,” dans *International conference on principles and practice of constraint programming*. Springer, 1998, p. 417–431.

[42] L. Musaraganyi *et al.*, “Integration of user’s preferences into the home healthcare routing and scheduling multi-objective problem : A hierarchical approach with pareto-optimal alternative solutions,” dans *International Conference on Human-Centred Software Engineering*. Springer, 2019, p. 179–191.

ANNEXE A VARIABLES EXPLICATIVES UTILISÉES

Tableau A.1 Variables explicatives qui proviennent des données d'AlayaCare

Variable	Explication	CAN	USA	AUS
n_visits	Nombre de visites sur une période	x	x	x
h_pv	Durée moyenne des visites	x	x	x
h_std	Écart-type des durées des visites	x	x	x
late_pv	Retard moyen sur une période	x	x	x
perc_clients	Rapport du nombre de clients distincts divisé par le nombre de visites	x	x	x
perc_servcodes	Rapport du nombre de services distincts divisé par le nombre de visites	x	x	x
n_cancelled	Nombre de visites annulées	x	x	x
est_dist_per_day	Distance moyenne parcourue chaque jour travaillé	x	x	x
weekend_percentage	Visites en fin de semaine	x	x	x
night_percentage	Visites de nuit	x	x	x
evening_percentage	Visites en fin de journée	x	x	x
peak_percentage	Visites aux heures de pointes	x	x	x
mean_days_between_shifts	Nombre de jours moyen entre deux jours travaillés	x	x	x
n_blacklist	Nombres d'entrées de l'employé dans les listes noires	x	x	x
h_availability	Nombre d'heures où l'employé a indiqué être disponible	x	x	x
h_unavailability	Nombre d'heures où l'employé a indiqué ne pas être disponible	x	x	x
age	Âge de l'employé	x	x	x
is_medical	Variable binaire si le métier de l'employé nécessite un haut niveau d'étude	x	x	x
is_male	Variable qui indique si l'employé est un homme	x	x	x

Tableau A.2 Variables explicatives qui proviennent des données de recensement

Variable	Explication	CAN	USA	AUS
population	Nombre d'habitants	x	x	x
irsad	*Index of Relative Socio-economic Advantage and Disadvantage*			x
perc_married	Pourcentage de personnes mariées	x		
income_median	Revenu médian	x		
income_perc_gov	Pourcentage des revenus provenant du gouvernement	x		
perc_income_avg_mdn	Rapport entre le revenu moyen et le revenu médian	x		
perc_mtongue_official	Pourcentage d'habitants qui parlent une langue maternelle officielle	x		
perc_citizen_can	Pourcentage de citoyens canadiens parmi la population	x		
perc_diplom	Pourcentage de personnes ayant un diplôme	x		
perc_bac_or_above	Pourcentage de personnes ayant un baccalauréat, une maîtrise ou un doctorat	x		
perc_transp_car	Pourcentage d'utilisation de la voiture dans les déplacements	x		
perc_transp_public	Pourcentage d'utilisation des transports publics dans les déplacements	x		