

Mémoire présenté la semaine du 09/01/2023
pour l'obtention du diplôme de
Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : **Emilien Chupot**

Titre : **Étude de la corrélation entre consommations
santé, variables macro et absentéisme**

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

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

Membres présents du jury de la filière

Olivier Lopez

Entreprise : AXA France 

Nom : Fabienne Cazals

Signature :



*Membres présents du jury de l'Institut
des Actuaires*

Directeur du mémoire en entreprise :

Nom : Fabienne Cazals

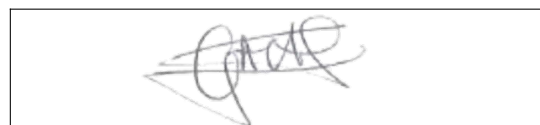
Signature :



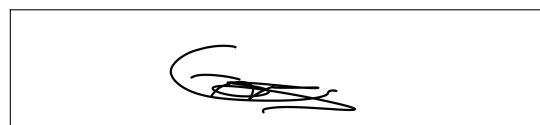
**Autorisation de publication
et de mise en ligne sur
un site de diffusion de
documents actuariels**

*(après expiration de l'éventuel délai
de confidentialité)*

Signature du responsable entreprise



Signature du candidat



Résumé

En 2020, la crise sanitaire liée à la Covid-19 est venue bouleverser la vie des entreprises et de leurs salariés. Elles ont dû faire face à des défis économiques et sociaux sévères et persistants. Cette période atypique que nous avons traversé a eu des conséquences lourdes sur la santé physique et psychologique des Français, ce qui est non sans impacter la productivité au travail (étude des Nations Unies [1]). Les Directions des Ressources Humaines doivent donc composer avec une part de Risques Psychosociaux grandissante, au-delà des aléas du quotidien que l'on regroupe autour des Troubles Musculo-squelettiques et qu'ils se doivent de maîtriser pour réduire un absentéisme très coûteux.

Ces risques émergents sont donc une opportunité pour les assureurs collectifs car ils ont l'expertise pour proposer une étude solide et une prévention au plus proche du terrain. La mise en application des Déclarations Sociales Nominatives (DSN) pour les entreprises du territoire national a considérablement changé la donne pour les opérateurs les recevant, les assureurs collectifs notamment. Sont concernées dans ces déclarations : les informations des systèmes de paie au sens large, et les signalements d'événements pour l'Assurance Maladie (arrêts maladies, accidents du travail, maternité, etc.), et ce pour chaque salarié. Leur caractère obligatoire à partir de 2017 permet d'avoir un panorama exhaustif du monde professionnel [2].

Enfin, la connaissance des consommations santé (consultation chez un médecin, achat de médicaments, passage à l'hôpital, séances chez le psychologue) est un formidable indicateur de risque pour la survenance d'un arrêt de travail.

De ce fait, ce mémoire propose une application réalisée sur un portefeuille d'actes auprès des salariés dont l'entreprise est couverte par **AXA France** en santé et prévoyance collectives, portefeuille enrichi par la DSN avec de précieux détails sur leur profil (caractéristiques individuelles, emploi, fréquence et motifs d'arrêts), et par de nombreuses variables macro créées en fonction de la date ou du lieu de survenance.

Mots-clé : Arrêt de travail, Covid, DSN, Machine Learning, régression multinomiale, prévention, engagement.

Abstract

In 2020, the Covid-19 crisis disrupted the operation of companies and of their employees. They had to face severe and persistent economic and social challenges. This atypical period that we have been through has had heavy consequences on the physical and psychological health of the French, which is not without impact on productivity at work (United Nations study [?, 1]. Human Resources Departments must therefore deal with a growing number of Psychosocial Risks, beyond the everyday hazards that are grouped around Musculoskeletal Disorders, which they must control in order to reduce very costly absenteeism. These emerging risks are therefore an opportunity for group insurers, as they have the expertise to propose a solid study and prevention as close to the field as possible. The implementation of Nominal Social Declarations (Déclarations Sociales Nominatives / DSN) for companies in France has considerably changed the situation for operators receiving them, notably group insurers. These declarations include information from payroll systems in the broadest sense of the term, as well as event reports for the health insurance system (sick leave, work accidents, maternity, etc.), for each employee. Their compulsory nature from 2017 onwards provides an exhaustive panorama of the professional world. Finally, knowledge of health consumption (visits to a doctor, purchase of medication, hospital visits, sessions with a psychologist) is a formidable risk indicator for the occurrence of a work stoppage [2].

This work proposes an application carried out on a portfolio of acts among employees whose company is covered by **AXA France** in group insurance, a portfolio enriched by the DSN with precious details on their profile (individual characteristics, employment, frequency and reasons for work stoppages), and by numerous macro variables created according to the date or place of occurrence.

Keywords : Interruption of work, Covid, DSN, Machine Learning, multinomial regression, prevention, involvement.

Note de synthèse

L'**absentéisme** est un phénomène qui coûte chaque année plus de **100 milliards d'euros** aux entreprises françaises. Plus inquiétant, c'est un phénomène qui prend de l'ampleur chaque année.

La crise sanitaire liée au Covid-19 est venue non seulement accentuer cet effet, mais a redessiné les contours de cette problématique. Cette crise est venue changer le rapport au travail des salariés. L'absentéisme est maintenant multifactoriel, avec les causes que l'on retrouvait avant la crise, mais on assiste à une forte croissance des troubles mentaux, d'où l'intégration du moral des ménages INSEE dans notre étude, ou encore l'avènement du télétravail ce qui peut éventuellement permettre d'expliquer une certaine dynamique d'arrêts de travail.

L'objectif de notre mémoire est donc de trouver des variables discriminantes parmi les consommations santé, ainsi qu'un panel de variables macro pour expliquer la survenance (ou non) des arrêts de travail, estimer leur durée et imaginer la stratégie du groupe pour anticiper ce risque.

Le périmètre concerné regroupe les arrêts de travail renseignés dans les Déclarations Sociales Nominatives survenus entre janvier 2019 et avril 2022, ainsi que les consommations santé des salariés couverts par AXA entre janvier 2017 et avril 2022 pour avoir un historique convenable.

Dans un premier temps, nous avons accompli un minutieux travail d'épuration de la table de travail. Nous avons retiré les montants de consommation santé aberrants, sélectionné les arrêts de travail appartenant spécifiquement à notre périmètre d'étude, éliminé les doublons ou anomalies. Nous avons ensuite rajouté des éléments macro permettant d'expliquer la survenance de l'arrêt de travail comme : le moral des ménages INSEE, la distance domicile-travail, la présence en zone embouteillée, la distance temporelle à des jours vaqués (vacances scolaires, jours fériés) ...

Ensuite, nous avons fusionné nos tables des DSN (activité des salariés) avec nos

tables de consommations santé. Nous avons agrégé ces dernières pour les synthétiser vis-à-vis de l'arrêt de travail (ou non) auquel elles sont potentiellement rattachées.

A ce stade, nous avons notre table d'étude ; nous pouvons lancer les modèles. Nous avons choisi de lancer 2 modèles successifs :

1. un modèle pour prédire la survenance (ou non) de l'arrêt de travail, grâce aux variables explicatives présentes et créées
2. un modèle pour prédire la durée de l'arrêt de travail sur la partie de la table prédite positivement en sortie du modèle 1., grâce aux variables explicatives présentes et créées

Modèle pour prédire la survenance ou l'absence d'arrêt de travail

Afin d'équilibrer notre table qui comporte environ 3/4 de lignes avec arrêt de travail pour 1/4 de lignes sans arrêt de travail, nous allons pondérer chaque ligne de la première cohorte avec un poids compris entre 0 et 1, proportionnel avec la durée de l'arrêt de travail concerné sur la durée totale d'arrêt de ce salarié sur son contrat de travail.

Notre sortie ici est binaire, 0 pour une absence d'arrêt de travail et 1 pour un arrêt de travail. Nous avons donc essayé plusieurs modèles :

1. plusieurs régressions logistiques
2. un modèle de forêts aléatoires / Random Forest
3. un modèle eXtreme Gradient Boosting / XGBoost

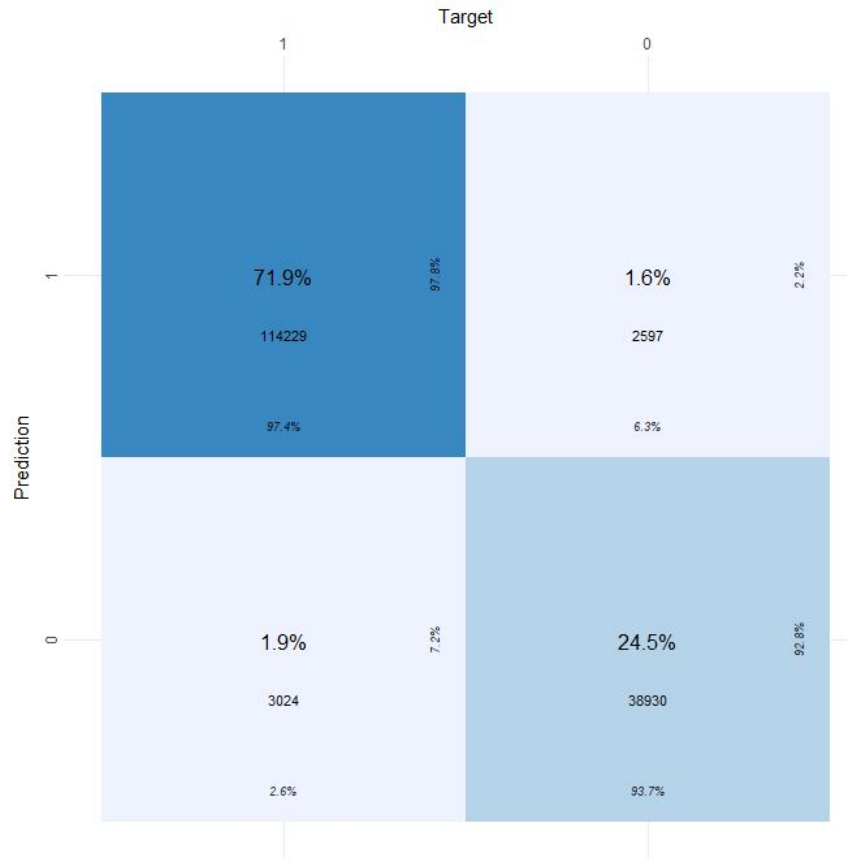
Voici les résultats de cette comparaison, et les caractéristiques du modèle retenu :

Modèle	R^2	Sensibilité	Spécificité
Régression Logistique sur toutes les var. (#1)	0,829	86,5%	72,8%
Régression Logistique améliorée (#2)	0,864	97,7%	54,5%
Random Forest	0,899	98,5%	65,4%
XGBoost	0,897	94,4%	76,7%
Régression Logistique #2 puis Random Forest	0,965	97,4%	93,7%
Régression Logistique #2 puis XGBoost	0,900	93,3%	80,7%

Le modèle retenu consistera alors en un enchaînement d'une régression logistique en écartant les variables non-significatives, et un Random Forest sur cette première prédiction. En effet, le premier modèle a une excellente sensibilité (*capacité à discerner les vrais positifs*) et le second a une excellente sensibilité + une spécificité très convenable (*capacité à trouver les vrais positifs et vrais négatifs, même dans un échantillon*

ne comportant *a priori* que des éléments positifs).

De cette manière, les lignes positives deux fois seront validées comme positives sinon les lignes négatives au moins une fois seront considérées comme négatives. Ci-dessous la matrice de confusion de ce double-modèle :



Matrice de confusion Régression Logistique + Random Forest pour prédire la survenance d'un arrêt de travail

Modèle pour prédire la durée des arrêts de travail

Pour ces travaux, nous allons partir de la base formée par les « 1 » en sortie du double-modèle précédent. L'objectif sera de prédire la durée de l'arrêt sous forme d'une variable continue sur N. Cette étude sera scindée en 2 études :

1. prédiction de la durée des arrêts de moins de 30 jours
2. prédiction de la durée des arrêts de plus de 30 jours

Au vu de la répartition de notre table, ce choix fait du sens (85% des arrêts durent moins de 30 jours), et cela permet de gagner en précision.

Modèle pour prédire la durée des arrêts de travail de moins de 30 jours

Cette première étude de notre second volet aura pour but de modéliser la durée des arrêts « courts ». Nous avons songé à utiliser plusieurs modèles linéaires généralisés (GLM) :

- GLM Poisson
- GLM Gamma
- Random Forest
- XGBoost

Tous ces modèles ont proposé peu ou prou des résultats similaires. Nous avons alors calculé une table de maintien en incapacité, ventilée selon plusieurs variables : genre, CSP, type de contrat, distance temporelle par rapport aux vacances scolaires, distance temporelle par rapport au week-end situation familiale, présence dans une zone embouteillée. Pour chaque colonne $k \in \llbracket 1, 30 \rrbracket$, on remplit le nombre de personnes l_k qui se maintenaient encore en arrêt au k -ème jour.

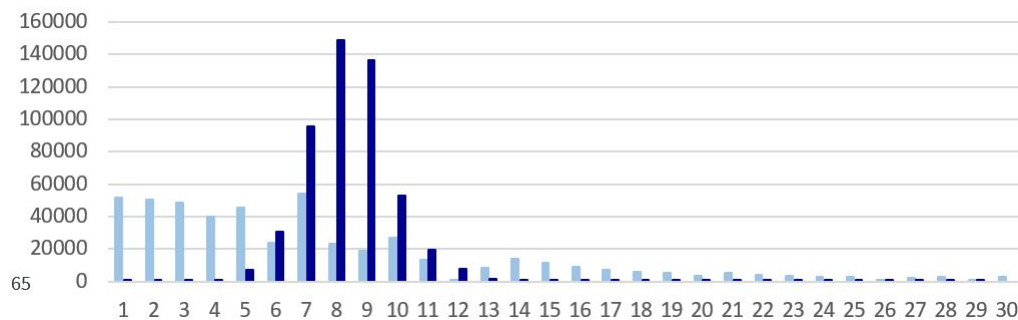
L'espérance de maintien de la ligne i se calcule en sommant les probabilités de maintien du jour 0 au jour k , c'est-à-dire en divisant par l_0 :

$$E_{\text{maintien},i} = e_{0,i} = \sum_{k=1}^{30} {}_k p_{0,i} = \frac{1}{l_{0,i}} \sum_{k=1}^{30} l_{k,i}$$

Ce qui donne (valeurs fictives) :

i	GENRE	TR AGE	CSP	Contrat	GdeVille	WE	Sit familiale	JV	0	1	...	30	E
1	F	20-24	Cadre	CDI	Oui	2j	En couple	0-9j	39	0,92	...	0	7
2	F	20-24	Cadre	CDI	Oui	2j	En couple	10-19j	25	0,88	...	0	6
3	F	20-24	Cadre	CDI	Oui	2j	En couple	20-29j	14	1	...	0	10
4	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	0-9j	35	0.91	...	0	5
5	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	10-19j	24	0,92	...	0	6
6	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	20-29j	16	0.88	...	0	7
...
fin	H	60-64	N.Cadre	CDD	Non	3j	Seul(e)	20-29j	26	0.85	...	0	3

Ce modèle si on le rapatrie sur notre base, nous donne les prédictions suivantes :



Répartition des durées d'arrêt de travail réelles (bleu clair) et prédites par la table de maintien en incapacité (bleu foncé)

La prédiction de la durée n'est pas excellente, et donne la même allure qu'avec des méthodes type GLM et *Machine Learning* : nous sommes confrontés à une distribution normale autour de 7 jours. Cependant, nous pouvons évaluer la pertinence du modèle en se plaçant à la maille du SIREN et du SIRET.

Voici les moyennes des résidus moyens par SIREN et SIRET pour nos (meilleurs) modèles :

	GLM Poisson	GLM Gamma	Table de maintien
Résidu moyen par SIREN	0,72j	0,74j	0,47j
Résidu moyen par SIRET	0,75j	0,68j	0,38j

Nous validons alors le choix de la modélisation de la durée de l'arrêt de travail à l'aide de cette table de maintien.

Modèle pour prédire la durée des arrêts de travail de plus de 30 jours

Pour les arrêts de plus de 30 jours s'est posée la question de regrouper les durées d'arrêt par classes : *1-3 jours*, *4-7 jours*, *8-30 jours*, *1-2 mois*, *2-3 mois*, *3-6 mois*, *6+ mois*. Le résultat n'était pas exploitable à cause d'une trop grand déséquilibre. Nous sommes donc restés sur une prédiction continue entre 30 jours et 3 ans.

Tous les modèles ont été essayés : GLM Poisson, GLM Gamma, GLM Binomial Négatif, Loi de Pareto Généralisée, Random Forest, XGBoost. Ces modèles ne se sont pas avérés être efficaces.

Nous avons donc songé à re-faire une table de maintien en arrêt de travail, mois par mois mais cette fois-ci non-ventilée. Cette table est donc construite sur la population générale, seulement étalée sur les âges de 20 à 65 ans. La ventilation arrive après en calculant des ***Standardized Mortality Ratio*** (SMR) pour le Genre, la CSP, le type de contrat et la situation familiale :

$$SMR_i = \frac{\text{Durée d'arrêt totale prédite pour la cohorte } i}{\text{Durée d'arrêt totale réelle pour la cohorte } i}$$

Nous obtenons les taux SMR suivants :

Variable	Modalité	Taux SMR
Genre	Homme	1,022
	Femme	0,973
CSP	Cadre	1,066
	Non cadre	0,99
Contrat	CDI	0,996
	CDD	1,332
Situation familiale	En couple	1,025
	Seul(e)	0,976

Ces coefficients viennent se greffer (en division) à la durée prédite pour l'âge du salarié, en fonction de ses caractéristiques individuelles.

Si l'on rapatrie toutes les prédictions à notre table initiale pour confronter le prédit et le réel, on obtient des résidus moyens par tête :

	Maintien arrêts courts	Maintien arrêts longs	Global
Résidu moyen par SIREN	0,47j	-8,84j	-1,61j
Résidu moyen par SIRET	0,75j	-9,39j	-1,8j

Ce modèle se trompe environ de 1,6 jours pour chaque salarié. Cela n'est pas très satisfaisant mais les données étaient bien trop dispersées pour avoir une modélisation exacte et précise.

De plus, les consommations santé n'intervenaient pas dans l'évaluation de la durée d'arrêt car tous les profils de consommations aboutissaient à toutes les durées d'arrêt possibles...

Dans quelle mesure est-il possible de faire baisser l'absentéisme ?

Dans ce chapitre est abordée la stratégie à adopter par l'assureur et par l'employeur pour tenter de réduire cet absentéisme au regard des contributions que l'on a pu observer dans nos modélisations précédentes, des tendances globales et des problèmes et solutions émergents.

Après avoir discuté des causes de l'absentéisme, il ressort deux axes d'action :

1. la nécessité de réduire l'impact de la maladie sur le travail, ce qui passe par la prévention fréquente, le traitement précoce et l'accompagnement des personnes fragiles
2. l'amélioration conséquente des conditions de travail pour obtenir l'engagement du salarié

Un plan d'action a été étudié dans ce mémoire, sur 3 entreprises du portefeuille :

	Entreprise A	Entreprise B	Entreprise C	Global
Caractéristique	plus fort taux d'absentéisme	plus forte démographie	plus fort chiffre d'affaires	-
Nb employés	3 200	13 000	1 150	250 000
CCN	Hôtellerie, Restauration, Tourisme	Bureau d'études, prestation de services aux entreprises	Banques, Étab. financiers, Assurances	-
Taux d'abs.	7,58%	2,20%	2,18%	3,69%
Cadres/NC	4/96%	92/8%	20/80%	17/83%

Ci-dessous les étapes appliquées dans le plan :

1. **Au titre de la prévention des Affections Longue Durée** : retirer 20% des arrêts au hasard si l'une des conditions est remplie :
 - le montant de consommation en examens est supérieur au quantile d'ordre 0,8
 - le montant de consommation en hospitalisation est supérieur au quantile d'ordre 0,8 et l'arrêt dépasse 1 mois
2. **Au titre de la santé psychologique** : retirer 21% des arrêts au hasard avec une consultation chez psychologue
3. **Au titre du télétravail** : retirer 15% des arrêts au hasard si le profil est un cadre habitant en zone dense et travaillant à + de 25 kilomètres de son domicile
4. **Au titre de l'amélioration des conditions de travail en général** : retirer 6% des arrêts au hasard sur la table restante

Chacune des étapes ci-dessus générera un périmètre sur lequel le taux d'absentéisme sera calculé. Cet algorithme sera répété **250 fois** puis la moyenne sera prise pour obtenir des valeurs stables et non soumises à des tirages qui risqueraient de biaiser l'étude.

Voici les impacts de chaque étape :

Évol. taux d'abs (%)	Entreprise A	Entreprise B	Entreprise C	Global
Impact Étape 1	-10	-11,7	-13,5	-11,5
Impact Étape 2	-1,5	-3,7	-3,2	-1,5
Impact Étape 3	-0,3	-2,7	-1,1	0
Impact Étape 4	-6,3	-6,3	-6,4	-6,5
Impact Total	-19	-26,4	-26	-20,6

Ce plan a des impacts différents en fonction de la démographie de l'entreprise concernée. Il faut donc adapter la stratégie, par exemple pour des entreprises comptant davantage de cadres : renforcer le suivi psychologique et proposer davantage de flexibilité. Pour une entreprise comptant des métiers plus physiques (vente, restauration, transports) : renforcer les bilans de santé, car les prix sont souvent élevés et le faible poste de dépense en examens cache peut-être un manque de soins.

Executive summary

Absenteeism is a phenomenon that costs French companies more than **100 billion euros** every year. More worrying, it is a phenomenon that is growing every year.

The health crisis linked to Covid-19 has not only accentuated this effect, but has also redrawn the outlines of this problem. This crisis has changed the employees' relationship to work. Absenteeism is now multifactorial, with the same causes as before the crisis, but there has been a strong increase in mental disorders, hence the inclusion of the INSEE household mood survey in our study, or the advent of teleworking, which may help explain a certain dynamic of work interruptions.

The objective of our paper is therefore to find discriminating variables among the health consumptions, as well as a panel of macro variables to explain the occurrence (or not) of work interruptions, to estimate their duration and to imagine the strategy of the group to anticipate this risk. scope of the study includes the work interruptions reported in the Nominal Social Declarations that occurred between January 2019 and April 2022, as well as the health consumption of employees covered by AXA between January 2017 and April 2022 in order to have a suitable history.

As a first step, we did a thorough job of cleaning up the work table. We removed the outlier health consumption amounts, selected the work interruptions specifically belonging to our scope of study, eliminated duplicates or anomalies. We then added macro elements to explain the occurrence of the work interruption such as : INSEE household mood, distance from home to work, presence in a congested area, temporal distance to days off (school vacations, public holidays)...

Then, we merged our DSN tables (employees' activity) with our health consumption tables. We aggregated them to synthesize them with respect to the interruption (or not) to which they are potentially attached.

At this stage, we have our study table ; we can launch the models. We have chosen to run 2 successive models :

1. a model to predict the occurrence (or not) of the interruption, thanks to the present explanatory variables and
2. a model to predict the duration of the interruption on the part of the table predicted positively at the output of model 1. thanks to the present and created explanatory variables

Model to predict the occurrence or absence of work interruption

In order to balance our table, which has about 3/4 of rows with work stoppages for 1/4 of rows without work stoppages, we will weight each row of the first cohort with a weight between 0 and 1, proportional with the duration of the work interruption concerned on the total duration of the work interruption of this employee on his work contract.

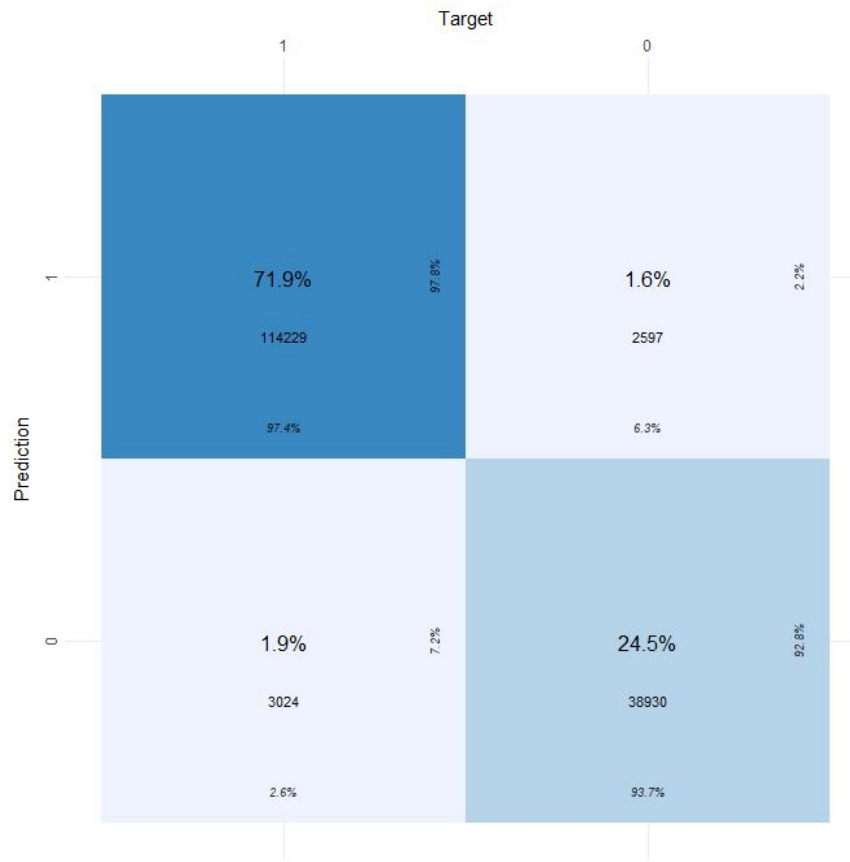
Our output here is binary, 0 for an absence of work interruption and 1 for a work interruption. So we tried several patterns :

1. several logistic regressions
2. a Random Forest model
3. an eXtreme Gradient Boosting / XGBoost model

Here are the results of this comparison, and the characteristics of the selected model :

Model	R^2	Sensibility	Specificity
Logistic Regression over every variable (#1)	0,829	86,5%	72,8%
Enhanced Logistic Regression (#2)	0,864	97,7%	54,5%
Random Forest	0,899	98,5%	65,4%
XGBoost	0,897	94,4%	76,7%
Logistic Regression #2 then Random Forest	0,965	97,4%	93,7%
Logistic Regression #2 then XGBoost	0,900	93,3%	80,7%

The selected model will then consist of a logistic regression by discarding the non-significant variables, and a Random Forest on this first prediction. Indeed, the first model has an excellent sensitivity (*ability to discern the true positives*) and the second one has an excellent sensitivity + a very suitable specificity (*ability to find the true positives and true negatives, even in a sample containing a priori only positive elements*). In this way, the positive lines twice will be validated as positive otherwise the negative lines at least once will be considered as negative. Below is the confusion matrix of this double model :



Confusion matrix of Logistic Regression + Random Forest to predict the predict the occurrence of work interruption

Model for predicting the duration of work interruptions

For this work, we will start from the base formed by the 1 « 1 » in output of the previous double-model. The objective will be to predict the duration of the interruption as a continuous variable on N. This study will be split into 2 studies :

1. prediction of the duration of interruptions of less than 30 days of the duration of interruptions of more than 30 days

In view of the distribution of our table, this choice makes sense (85

Model for predicting the duration of work interruptions of less than 30 days

This first study of our second component will aim to model the duration of « short » interruptions. We have considered using several generalized linear models (GLM) :

- GLM Poisson
- GLM Gamma
- Random Forest

— XGBoost

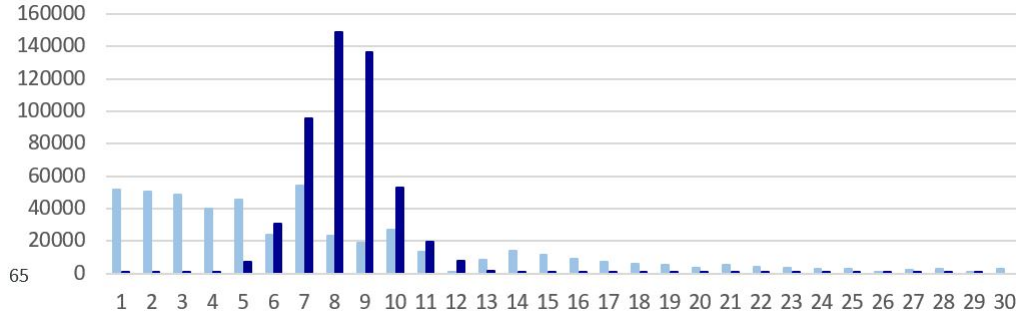
All these models suggested more or less similar results. We then computed a table of disability retention, differentiated according to several variables : gender, SPC, type of contract, distance from school vacations, distance from weekends, family situation, presence in a congested area. For each column $k \in \llbracket 1, 30 \rrbracket$, we fill in the number of people l_k who were still in interruption on the k -th day. holding expectation of line i is calculated by summing the holding probabilities from day 0 to day k , i.e. by dividing by l_0 :

$$E_{retention,i} = e_{0,i} = \sum_{k=1}^{30} k p_{0,i} = \frac{1}{l_{0,i}} \sum_{k=1}^{30} l_{k,i}$$

Which gives (imaginary values) :

i	GENDER	AGE	SPC	Contract	City	WE	Family Sit*	Holidays	0	1	...	30	E
1	F	20-24	Executive	Permanent	Yes	2d	w/ someone	0-9d	39	0,92	...	0	7
2	F	20-24	Executive	Permanent	Yes	2d	w/ someone	10-19d	25	0,88	...	0	6
3	F	20-24	Executive	Permanent	Yes	2d	w/ someone	20-29d	14	1	...	0	10
4	F	20-24	Executive	Permanent	Yes	2d	w/ someone	0-9d	35	0,91	...	0	5
5	F	20-24	Executive	Permanent	Yes	2d	w/ someone	10-19d	24	0,92	...	0	6
6	F	20-24	Executive	Permanent	Yes	2d	w/ someone	20-29d	16	0,88	...	0	7
...
end	H	60-64	N_Executive	Fixed	No	3d	Single	20-29d	26	0,85	...	0	3

This model if repatriated to our base, gives us the following predictions :



Distribution of actual (light blue) and predicted (dark blue) duration of work interruptions

The prediction of the duration is not excellent, and gives the same appearance as with methods such as GLM and *Machine Learning* : we are faced with a normal distribution around 7 days. However, we can evaluate the relevance of the model by looking at the SIREN and SIRET grid.

Here are the average residuals by SIREN and SIRET for our (best) models :

	GLM Poisson	GLM Gamma	Retention table
Average residual per SIREN	0,72d	0,74d	0,47d
Average residual per SIRET	0,75d	0,68d	0,38d

We then validate the choice of modeling the duration of the interruption using this holding table.

Model to predict the duration of work interruptions of more than 30 days

For interruptions longer than 30 days, the question arose of grouping the duration of the interruption by class : *1-3 days, 4-7 days, 8-30 days, 1-2 months, 2-3 months, 3-6 months, 6+ months* The result was not usable because of a too big imbalance. So we stayed on a continuous prediction between 30 days and 3 years.

All the models were tried : GLM Poisson, GLM Gamma, GLM Binomial Negative, Generalized Pareto Law, Random Forest, XGBoost. These models did not prove to be efficient.

We therefore thought of re-doing a table of maintenance in interruption of work, month by month but this time not ventilated. This table is therefore built on the general population, only spread over the ages from 20 to 65 years. The breakdown comes afterwards by calculating the Standardized Mortality Ratio (SMR) for gender, social class, type of contract and family situation :

$$SMR_i = \frac{\text{Total predicted interruption duration for the cohort } i}{\text{Total actual interruption duration for the cohort } i}$$

We get the following SMR :

Variable	Modality	SMR
Gender	Male	1,022
	Female	0,973
SPC	Executive	1,066
	Non-executive	0,99
Contract	Permanent	0,996
	Fixed	1,332
Family Situation	w/ someone	1,025
	Single	0,976

These coefficients are superimposed (with a division) on the predicted duration for the employee's age, depending on his individual characteristics.

If we bring back all the predictions to our initial table to compare the predicted and the real, we obtain average residuals per head :

	Retention short int.	Retention long int.	Global
Average residual per SIREN	0,47d	-8,84d	-1,61d
Average residual per SIRET	0,75d	-9,39d	-1,8d

This model is wrong by about 1.6 days for each employee. This is not very satisfactory, but the data were far too scattered to have an accurate and precise model. Moreover, the health consumptions did not intervene in the evaluation of the duration of interruption because all the consumption profiles resulted in all the possible durations of interruption...

To what extent is it possible to reduce absenteeism ?

This chapter discusses the strategy to be adopted by the insurer and the employer in an attempt to reduce absenteeism in the light of the contributions observed in our previous modelling, global trends and emerging problems and solutions.

After discussing the causes of absenteeism, two lines of action emerge :

1. the need to reduce the impact of illness on work, which requires frequent prevention, early treatment and support for fragile people
2. the consequent improvement of the working conditions to obtain the commitment of the employee

An action plan has been studied in this thesis, on 3 companies of the portfolio :

	Company A	Company B	Company C	Global
Feature	highest absenteeism rate	highest demography	highest turnover	-
Nb employees	3,200	13,000	1,150	250,000
NCC	Hotel, Restaurant, Tourism Industries	Design offices, services to companies	Banks, Financial institutions Insurers	-
Abs. rate	7,58%	2,20%	2,18%	3,69%
Exec/Non-exec	4/96%	92/8%	20/80%	17/83%

Below are the steps applied in the plan :

1. **For the prevention of long-term illnesses** : remove 20% of the interruptions at random if one of the conditions is met :
 - the amount of consumption in examinations is higher than the quantile of order 0.8
 - the amount of inpatient consumption is greater than the quantile of order 0.8 and the interruption exceeds 1 month
2. **For psychological health** : remove 21% of the interruptions at random with a psychologist consultation

3. **For remote working** : remove 15% of the interruptions at random if the profile is an executive living in a dense area and working more than 25 kilometers from his home
4. **For the improvement of working conditions in general** : remove 6% of the interruptions at random from the remaining table

Each of the above steps will generate a perimeter over which the absenteeism rate will be calculated. This algorithm will be repeated **250 times** and then the average will be taken to obtain stable values that are not subject to draws that could bias the study.

Here are the impacts of each step :

Var. abs. rate (%)	Company A	Company B	Company C	Global
Impact of Step 1	-10	-11,7	-13,5	-11,5
Impact of Step 2	-1,5	-3,7	-3,2	-1,5
Impact of Step 3	-0,3	-2,7	-1,1	0
Impact of Step 4	-6,3	-6,3	-6,4	-6,5
Global Impact	-19	-26,4	-26	-20,6

This plan has different impacts depending on the demographics of the company concerned. The strategy must therefore be adapted, for example, for companies with more executives : strengthen psychological follow-up and offer more flexibility. For a company with more physical occupations (sales, catering, transport) : increase health check-ups, as prices are often high and the low expenditure on examinations may hide a lack of care.

Table des matières

Résumé	1
Abstract	2
Note de synthèse	3
Executive summary	10
Introduction	20
Préliminaires : Le risque arrêt de travail en France - Conception de la	
table de travail	22
La prévoyance	22
Les contrats de prévoyance	23
Prévoyance individuelle	23
Prévoyance collective	23
Le risque arrêt de travail en France	24
L'état d'incapacité	25
Définition - Eligibilité	25
Les prestations versées par la Sécurité Sociale	25
Les prestations versées par l'entreprise	26
Les prestations versées par l'organisme complémentaire	26
Périmètre de l'étude	28
Environnement et outils de travail	28
Table des Actes en Santé	29
Tables des DSN	29
Conception de la table d'étude / Fusion des tables Santé et DSN	33
Critères de jointure entre les tables d'Actes Santé et DSN	33
Choix de la maille d'étude	33
Pondération des lignes	34
Agrégation des arrêts de travail antérieurs (profil d'inactivité)	36
Traitement des anomalies	36

Création d'indicateurs macro	40
Variables propres à la date de survenance de l'élément déclencheur j . .	41
Statistiques descriptives	44
Modélisation du risque incapacité	47
Prédiction de la survenance d'un arrêt de travail	47
Généralités sur les modèles linéaires	47
Théorie de la régression logistique	49
Modélisation de la survenance de l'arrêt de travail par régression logistique	49
Random Forest	52
XGBoost	53
Amélioration de la modélisation en combinant les méthodes	55
Régression Logistique puis Random Forest	55
Régression Logistique puis XGBoost	56
Choix de la modélisation de la survenance de l'arrêt de travail	56
Interprétation du modèle	58
Prédiction de la durée d'un arrêt de travail	62
Modélisation de la durée des arrêts de moins d'1 mois	64
Calcul des taux de maintien	70
Rapatriement des prédictions par la table de maintien en incapacité sur notre table d'étude	73
Modélisation de la durée des arrêts > 30 jours	74
Concaténation des 2 tables de maintien en arrêt « court » et « long » avec leurs prédictions	76
Comment maîtriser cet absentéisme au regard de cette étude ?	77
Généralités sur l'absentéisme	77
Les coûts de l'absentéisme	77
Les coûts directs de l'absentéisme	78
Les coûts indirects de l'absentéisme	78
Les coûts cachés de l'absentéisme	78
L'évolution de la situation ces dernières années	78
La prévention comme l'une des solutions	80
Améliorer les conditions de travail pour améliorer l'engagement	81
Proposition de plan de réduction de l'absentéisme et simulation de son impact sur le taux d'absentéisme	82
Présentation des entreprises témoin	83
Hypothèses	84
Hypothèse de prise en charge précoce des ALD	84
Hypothèse de prise en charge précoce des troubles psychologiques	85

Hypothèse d'amélioration des conditions de travail	85
Résultats de la modélisation	86
Conclusion	88
Remerciements	90
Annexes	93

Introduction

L’absentéisme est un phénomène coûtant chaque année plus de 100 milliards d’euros aux entreprises françaises, que ce soit en maintien de salaire, manque à gagner, ou en image [3]. Celui-ci suit une tendance inquiétante depuis la crise sanitaire avec la multiplication des arrêts pour cause de Covid, la volonté des salariés de trouver un meilleur équilibre vie professionnelle-vie personnelle et la généralisation du télétravail [4] [5].

C’est à ce niveau que l’assureur collectif peut jouer un rôle essentiel pour conseiller ses entreprises clientes : il a le savoir-faire actuariel et a à sa disposition une myriade de données que ce soit via les DSN et par les remboursements en santé collective. Ce secteur est donc porteur de par sa croissance et de par la préoccupation qu’il génère. AXA étant l’un des leaders sur le marché des collectives françaises, concentre une énergie grandissante à cerner les tendances de l’absentéisme dans le pays.

L’objectif de cette étude alors est de comprendre s’il y a des facteurs engendrant davantage de sinistralité en arrêt de travail, si ces variables peuvent nous indiquer si l’arrêt sera court ou long et dans quelle mesure **AXA** peut orienter sa stratégie pour réduire cet absentéisme. Cette fin est gagnante-gagnante-gagnante pour l’assureur, l’employeur (pour qui cela coûte cher) et l’employé (pour qui l’état de santé s’est dégradé).

Ce mémoire répond à une demande qui a été exacerbée à cause de la crise sanitaire. Cette étude se fait donc au juste lendemain de cette crise, et a partiellement été abordée soit avant soit pendant le Covid. De plus, le lien entre santé et absentéisme étant concret, l’apport des consommations de santé et de plusieurs variables macro permet d’enrichir l’étude. Aussi, au vu du grand format des données à disposition ainsi que de leur efficacité, des méthodes de *Machine Learning* seront privilégiées et confrontées à des méthodes plus *mainstream* en actuariat.

Ce mémoire est scindé en **3 parties** :

La première partie présente les enjeux et le point de départ de l'étude. Dans un premier temps, elle rappelle le contexte de la prévoyance collective, le fonctionnement du risque arrêt de travail en France et les remboursements prévus. Ensuite sont abordés les données à notre disposition et la conformité avec la réglementation en vigueur (RGPD). Dans un dernier temps sera évoquée en détail la création de la table ainsi que les critères et le périmètre retenus.

L'objet de la seconde partie est l'étude de nos données. Dans un premier temps sera modélisée la survenance de l'arrêt de travail, avec toutes les variables qui expliquent ce sinistre. Dans un second temps sera modélisée la durée des arrêts de travail survenus avec le même objectif de comprendre quelle variable impacte la durée de l'arrêt.

Enfin, dans la troisième partie seront tirées les conclusions de la partie précédente. Cette dernière partie a pour but d'analyser la situation actuelle au regard de nos modèles, de proposer des solutions, de simuler un plan pour réduire l'absentéisme en tirant des impacts chiffrés.

Contexte

La prévoyance

La prévoyance est née avec la loi n 89-1009 dite **loi EVIN** [6]. Elle est définie comme « les opérations ayant pour objet la prévention et la couverture du risque décès, des risques pourtant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque chômage ».

Elle désigne de ce fait dans le cadre de l'assurance les garanties et contrats qui couvrent l'assuré de 4 risques majeurs de la vie :

- **Incapacité de travail** : la garantie incapacité de travail permet selon les contrats, de conserver un certain niveau de revenu, voire la totalité. Ces indemnités dites journalières complémentaires viennent compléter les indemnités versées par la **Sécurité Sociale** et la part de salaire maintenue par l'employeur.
- **Invalidité** après 3 ans en incapacité ou si vous êtes reconnu partiellement/totalement invalide après un accident ou une maladie, cette garantie consiste en le versement d'une rente complémentaire à l'indemnisation versée par la **Sécurité Sociale** pour revenir à 100% du salaire.
- **Décès** : cette garantie prend la forme d'un **capital** ou d'une **rente**, versé(e) au conjoint ou aux enfants (rente d'éducation). Ce contrat peut aussi inclure d'autres garanties telles qu'un forfait pour les obsèques, une majoration en cas de décès simultané du conjoint ou un doublement du capital en cas de décès accidentel.
- **Dépendance** : cette garantie prévoit le versement d'une **rente viagère** ou d'un **capital**.

On distingue deux catégories de contrats de prévoyance :

- **la prévoyance individuelle**
- **la prévoyance collective**

Les contrats de prévoyance

Le **contrat de prévoyance** fait intervenir plusieurs acteurs : le souscripteur, les affiliés, les assurés et le(s) bénéficiaire(s). Le souscripteur correspond à la personne physique (individuel) ou morale (collectif) qui cotise. Les affiliés sont des personnes du groupe assurable : salariés d'une entreprise ou membres d'une association. Les assurés sont ceux sur qui repose le risque. Les bénéficiaires sont ceux qui sont susceptibles de percevoir les prestations en cas de réalisation du sinistre.

Prévoyance individuelle

La prévoyance individuelle concerne les contrats souscrits par l'assuré auprès d'un organisme d'assurance, sans intermédiaire. Celui-ci est donc le seul payeur de sa cotisation.

Ce type de contrat est donc plus adapté au profil de risque du souscripteur car il est établi uniquement à l'aide de ses caractéristiques individuelles.

L'adhésion à un contrat de prévoyance individuelle peut être soumise à une sélection médicale (questionnaires médicaux, examens spécifiques). L'organisme d'assurance peut ensuite faire appel à un médecin-conseil qui jugerait de l'acceptabilité de ce risque à assurer, et peut aussi conduire à une adaptation du tarif de la prime à payer pour l'assuré.

Enfin, chaque changement de l'état de santé de l'assuré devra être déclaré à l'assureur.

Prévoyance collective

La prévoyance collective [7] est proposée par certaines entreprises à leurs salariés afin de leur faire bénéficier d'une couverture complémentaire en terme de protection sociale. Souvent prévue dans le cadre d'accords collectifs, elle peut être instaurée à la suite d'un référendum ou par décision unilatérale de l'employeur.

Les cotisations d'un contrat collectif prévoient le partage des cotisations entre l'employeur et le salarié, ce qui permet de diminuer le tarif et revêt un caractère avantageux pour le salarié car il bénéficiera pleinement des garanties en cas de sinistre.

On parle de prévoyance collective **facultative**, qui se distingue de la prévoyance collective **obligatoire** : le choix est laissé libre aux salariés d'y adhérer ou non. De plus, l'entreprise peut prévoir des garanties optionnelles à **adhésion facultative** pour les salariés qui souhaiteraient bénéficier de garanties améliorées (dans les faits, c'est rarement le cas car les entreprises ne peuvent pas bénéficier d'exonérations de charges sociales).

Enfin, depuis le 1^{er} Juin 2015, les salariés bénéficient de ce que l'on appelle **la portabilité** [8] : ce sont des dispositifs qui permettent à un salarié après son départ de

l'entreprise et sous conditions de continuer à bénéficier des couvertures en vigueur chez son ex-employeur pendant un certain temps (souvent : **1 an**).

L'adhésion à ce contrat de prévoyance collective offre de nombreux avantages :

- une meilleure protection sociale complémentaire que sur un contrat d'assurance individuel
- les garanties s'appliquent à tous les assurés, quel que soit leur âge, revenu ou état de santé à la souscription
- des coûts réduits du fait de la mutualisation des risques
- les cotisations sont déductibles de l'impôt sur le revenu si l'adhésion est obligatoire
- ce contrat couvre également le conjoint

AXA France possède une part importante de contrats de prévoyance collective au sein de son portefeuille (chiffre d'affaires 2020 = 1 milliard d'euros).

Dans cette étude, nous nous intéresserons uniquement aux garanties « **incapacité de travail** » des contrats de prévoyance collective.

Le risque arrêt de travail en France

Le **risque arrêt de travail** correspond au risque pour le salarié de devoir interrompre partiellement / définitivement son activité professionnelle à la suite d'un accident ou d'une maladie.

Cela va donc engendrer une perte partielle / totale de ses revenus.

L'arrêt de travail peut être déclenché après un accident de la vie quotidienne, une maladie ou une hospitalisation, mais aussi un accident de travail, de trajet ou à cause d'une maladie professionnelle.

Nous devons distinguer 2 états d'arrêt de travail, selon la **Sécurité Sociale** :

- **l'incapacité** : le sinistre interrompt partiellement l'activité professionnelle et les revenus de l'assuré.
- **l'invalidité** : le sinistre a engendré une réduction de la capacité de travail d'au moins 2/3 et ce, de manière permanente.

L'état d'invalidité revêt donc une nature **irréversible**. De plus, ces deux états sont liés : au bout de 3 ans d'incapacité, l'individu redevient valide **ou** passe automatiquement en invalidité.

Comme notre étude portera sur l'état d'incapacité, nous restreindrons la durée des arrêts de travail à sélectionner seulement si leur durée est inférieure à **3 ans**. La partie suivante plantera le décor quant à cet état...

L'état d'incapacité

Définition - Eligibilité

Comme présenté au-dessus, l'état d'incapacité temporaire de travail (ou **ITT** ou état d'incapacité) fait référence pour l'assuré victime d'un accident ou d'une maladie à une impossibilité de poursuivre son travail.

Cet état n'est éligible que pour les salariés en activité, ils doivent donc être âgés d'au moins **18 ans** pour y souscrire, et les garanties prennent fin au départ en **retraite de l'assuré**, ou à **65 ans**. De plus, cet état est potentiellement de nature répétitive au cours de la vie active d'un salarié.

Les prestations versées par la Sécurité Sociale

La **Sécurité Sociale** correspond au premier niveau de couverture. Elle dispose d'un **Code** [9] dont les articles L.321-1 et L.433-1 viennent définir l'état d'incapacité temporaire de travail. Dès qu'elle est constatée par le médecin traitant, l'assuré perçoit des **indemnités journalières** (IJ) pendant 3 ans maximum, plus ou moins élevées en fonction du salaire de l'assuré et de la cause de l'arrêt de travail. De plus, les modalités d'indemnisation sont différentes en fonction de ces facteurs :

	Maladie	Accident du travail Maladie professionnelle	
		avant 28j	après 28j
Durée de l'arrêt	Toutes	avant 28j	après 28j
Délai de carence	3 jours	0 jour	
Fraction du salaire	50%	60%	80%
Plafond de l'IJ	49,68€	205,84€	274,46€

Dans le cas d'un arrêt d'une **durée inférieure à 6 mois**, l'assuré doit pouvoir justifier d' :

- avoir travaillé au moins 150 heures au cours des 3 mois civils ou 90 jours précédant l'arrêt
- avoir cotisé au cours des 6 mois civils précédant l'arrêt sur une base de rémunération $\geq 1\,015x$ le SMIC **horaire**

Dans le cas d'un arrêt d'une **durée supérieure à 6 mois**, l'assuré doit pouvoir justifier d' :

- une affiliation à un régime de Sécurité Sociale depuis 12 mois au moins

- avoir travaillé au moins 600 heures au cours des 12 mois civils ou 365 jours précédant l'arrêt
- avoir cotisé au cours des 12 mois civils précédant l'arrêt sur une base de rémunération $\geq 2\,030 \times$ le SMIC **horaire**

Ces conditions remplies, les opérateurs des **Caisses Primaires d'Assurance Maladie** (CPAM) et de la **Sécurité Sociale** peuvent effectuer un certain nombre de contrôles visant à détecter la fraude (ex : vérifier si le salarié est apte à reprendre une activité). Ces décisions peuvent mettre fin au versement des indemnités journalières, voire sanctionner le salarié.

Les prestations versées par l'entreprise

Le deuxième niveau de couverture correspond aux prestations versées par l'employeur imposées par la **loi de mensualisation** de 1978 et renforcées par l'**Accord National Interprofessionnel** de 2008.

L'employeur est tenu de compléter les IJ versées par la Sécurité Sociale si le salarié a une **ancienneté supérieure à 1 an**, à hauteur de **90% du salaire brut** puis **2/3 du salaire brut** pendant des durées variant en fonction de l'ancienneté du salarié dans son entreprise :

Ancienneté	90% du salaire brut	2/3 du salaire brut
1 à 6 ans	30j	30j
6 à 11 ans	40j	40j
11 à 16 ans	50j	50j
16 à 21 ans	60j	60j
21 à 26 ans	70j	70j
26 à 31 ans	80j	80j
> 31 ans	90j	90j

De plus, il y a une **franchise de 7 jours** à observer avant le versement du complément d'IJ par l'employeur en cas d'**arrêt maladie**.

Les prestations versées par l'organisme complémentaire

L'organisme complémentaire vient en 3^{ème} temps compléter les indemnités versées par la Sécurité Sociale et l'employeur afin de permettre au salarié de conserver un certain niveau de revenu voire l'intégralité de son salaire.

Sont habilités à verser cette troisième couche (selon la **loi Evin**) :

- **les compagnies d'assurance**, régies par le Code des Assurances
- **les mutuelles**, régies par le Code de la Mutualité

— **les instituts de prévoyance**, régis par le Code de la Sécurité Sociale et le Code rural

Ces organismes ne verseront les indemnités complémentaires qu’après avoir reconnu l’état d’**incapacité** du salarié.

Nous venons donc de voir dans cette partie le contexte qui a mené à ce mémoire, nous allons maintenant prendre connaissance du périmètre de l’étude dans le chapitre suivant.

Conception de la table de travail

Périmètre de l'étude

Un projet fonctionnant grâce à la **Data Science** se doit d'avoir des données d'excellente qualité. Il faut donc y accéder, s'assurer qu'elles soient lisibles et si besoin est, de traiter toutes les anomalies possibles par des considérations métier.

Dans un premier temps, nous allons parler des technologies utilisées. Ensuite, nous ferons l'inventaire des données utilisées ; leur source, leur définition, leurs interactions, leurs améliorations. Enfin seront présentées des Statistiques Descriptives de notre table.

Environnement et outils de travail

Les données à destination de l'étude sont des informations en lien avec les contrats complémentaires santé individuels **et** collectifs. Ces données sont difficiles d'accès du fait du **Règlement Général pour la Protection des Données**, car elles figurent en partie dans les **Déclarations Sociales Nominatives** (DSN).

Cependant, leur quantités sont suffisamment importantes pour être exploitées dans cette étude.

L'accès et l'exploration des données a été effectué sous **Spark** (qui supporte les langages Python, Scala, Java, et SQL). **Spark** est une infrastructure logicielle *open-source* de calcul distribué, qui permet d'accélérer grandement les traitements grâce à la mise en parallèle des tâches sur plusieurs machines, en plus de leur exécution en mémoire et en temps réel.

Cela permet de réduire drastiquement les temps de calcul sur des bases de données volumineuses.

Source des données

Les données brutes sont issues d'un Datalake. Elles sont retraitées par l'équipe **Data Santé & Collectives** d'AXA France et extraites pour former des datamarts. Ce sont donc des segmentations du Datalake qui sont plus pratiques et épurées.

En plus de ces datamarts, nous avons un système de gestion nommé « **Pegase** », qui concentre les informations en lien avec les contrats (souscription, prise en charge de l'assuré, consommation santé), l'assuré ou le professionnel de santé (tiers-payant).

Les tables que nous avons en notre possession sont certes vastes (et volumineuses !) mais sont une mine d'or. Nous avons une vision très précise de l'activité en France à l'aide de quelques retraitements.

Table des Actes en Santé

Cette table contient des renseignements sur les salariés couverts par un contrat de santé collective.

Y figurent notamment :

- leur identité (nom, prénom et numéro de Sécurité Sociale anonymisés), adresse, situation familiale
- leur entreprise (nom, lieu)
- leur consommation santé (montant, type, lieu, date, répartition du remboursement, professionnel de santé).

La maille de cette table est la consommation santé (1 ligne = 1 consommation santé).

Mettons cette table de côté pour l'instant, pour travailler sur nos DSN afin de faire une jointure en adéquation avec le besoin de l'étude...

Tables des DSN

Les Déclarations Sociales Nominatives (DSN) sont un formidable outil pour avoir une vision actualisée chaque mois du monde du travail.

Les DSN un système permettant à tout employeur de déclarer en ligne et de façon unique un grand nombre d'informations de ses salariés concernant leur activité : salaires, arrêts de travail, fin de contrat, reprise anticipée, etc. Ces DSN étant la dernière étape de la paie, elles sont fournies par un logiciel de paie compatible. Une DSN couvre un périmètre de 1 SIRET, ce qui signifie qu'une entreprise devra déclarer ces informations autant de fois qu'elle a d'établissements. Elles servent à centraliser toutes les déclarations aux organismes sociaux :

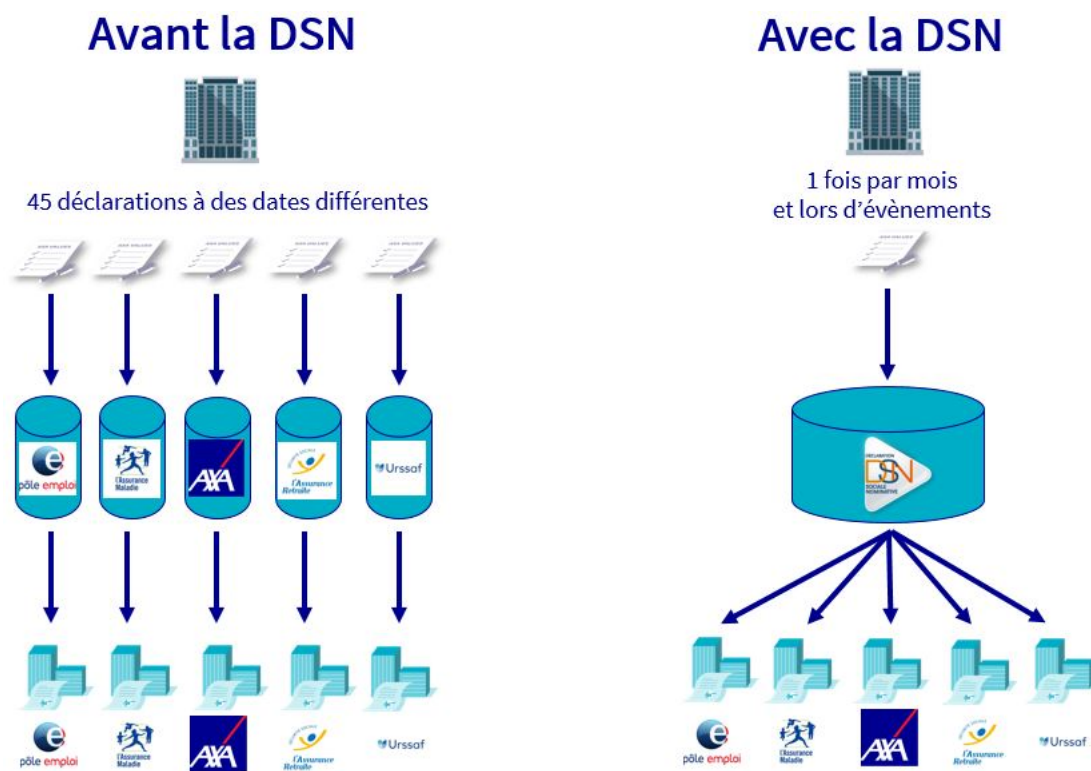


FIGURE 1 – Présentation des DSN

AXA France reçoit les données liées à son périmètre d'assureur, ie : Adhésion Prévoyance, "Affiliation Prévoyance", "Arrêt de travail". Les travaux que réalisent une partie de l'équipe **Data Santé & Collectives** consiste à construire des tables exploitables à partir de ces blocs de DSN, et nous aurons besoin notamment de 2 de ces tables pour avoir une vision cohérente et complète des salariés dont nous voulons étudier la consommation santé :

Table d'activité des DSN des entreprises couvertes par AXA

La table d'activité des DSN contient les informations sur les périodes d'activité des salariés couverts par AXA en collectif (Santé, Prévoyance, Dépendance, Retraite).

NB : La jointure avec la table des Actes en Santé ne nous fournira évidemment que les salariés couverts en Santé.

La maille de cette table est le contrat de travail unique : (1 ligne = 1 contrat de travail dans 1 entreprise pour 1 individu). La mission ici alors est de trouver le contrat de travail le plus pertinent pour notre étude pour chaque individu dans le cas où un individu ait plusieurs emplois. Les règles de sélection de ce contrat "cible" sont les suivantes :

- CDI > CDD
- Si doublons : mise à jour DSN la plus récente
- Si doublons : contrat le plus récent

- Si doublons : les contrats doublons coïncident en toute variable exploitable alors choix arbitraire (SIRET le plus faible)

La table d'activité a désormais un format similaire à :

NIR	Deb_Ctr_Tr	Fin_Ctr_Tr	Nom	Prénom	Adresse	Genre	...
NIR 1	22/04/1996	18/10/2011	NoM1	Prénom1	Adresse1	F	...
NIR 1	19/10/2011	31/12/2099	NoM1	Prénom1	Adresse1	F	...
NIR 2	29/08/2003	31/12/2099	NoM2	Prénom2	Adresse2	H	...
NIR 3	01/01/1971	30/11/2019	NoM3	Prénom3	Adresse3	H	...
NIR 4	30/04/1996	31/12/2099	NoM4	Prénom4	Adresse4	F	...
...

...	Nom_Entp	Adr_Entp	Type_Ctr_Tr	CSP	Motif_Rupture
...	Numéricable	Lyon	CDI	Cadre	changement d'emploi
...	SFR	Villeurbanne	CDI	Cadre	-
...	Carrefour	Nantes	CDD	Salarié	-
...	Renault	Achères	CDI	Ouvrier	Départ à la retraite
...	Sodexo	Paris 12ème	CDI	Salarié	-
...

NB : Les faux caractères signifient que les données ont été cryptées.

Table d'inactivité des DSN des entreprises couvertes par AXA

Cette table recense tous les arrêts de travail des individus de la table d'activité. On a donc pour chaque individu (et chaque contrat de travail qu'il possède) ses dates d'arrêt et certaines de ses informations personnelles. Nous écarterons par principe les arrêts pour maternité, paternité et adoption. Ces arrêts ne représentent pas un absentéisme que nous cherchons à éliminer (ils sont peu ou prou indépendants du jour de l'année ou des consommations en santé).

Cette table, assez légère en retraitements, aura cette allure :

NIR	Deb_AT	Fin_AT	Durée_AT	Nom	Prénom	Genre	Motif_AT	...
NIR 2	11/02/2019	17/02/2019	7	NoM2	Prénom2	H	Maladie	...
NIR 2	28/07/2021	30/07/2021	3	NoM2	Prénom2	H	Maladie	...
NIR 3	10/01/2017	18/02/2019	771	NoM3	Prénom3	H	Accident travail	...
NIR 4	20/05/2020	24/05/2020	5	NoM4	Prénom4	F	Maladie	...
...

Fusion des Tables des DSN

Le but de cette étape est de créer une version améliorée de l'activité (et par extension inactivité) provenant de nos DSN. Puisque nous voulons prédire la survenance **ou non** de l'arrêt de travail, nous devons joindre :

Activité *LEFT JOIN* Inactivité

(enrichir l'inactivité par les infos de l'activité)

La jointure se fera sur la clé commune aux 2 tables, *i.e.* **le NIR**, mais nous devons aussi imposer une condition supplémentaire pour que une partie de la jointure (l'interne) soit sensée :

Début du contrat de travail \leq Début de l'arrêt de travail \leq Fin du contrat de travail

Cette condition est essentielle pour rattacher les bonnes informations professionnelles au bon arrêt de travail, et *vice versa*.

À cette étape nous avons une table recensant tous nos contrats de travail de salariés, avec leurs arrêts de travail **ou non**. Elle a un format similaire à :

NIR	Deb_Ctr_Tr	Fin_Ctr_Tr	Nom	Prénom	Adresse	Genre	Nom_Entp	...
NIR 1	22/04/1996	18/10/2011	NoM1	Pr€n0m1	Adr€s\$e1	F	Numéricable	...
NIR 1	19/10/2011	31/12/2099	NoM1	Pr€n0m1	Adr€s\$e1	F	SFR	...
NIR 2	29/08/2003	31/12/2099	NoM2	Pr€n0m2	Adr€s\$e2	H	Carrefour	...
NIR 2	29/08/2003	31/12/2099	NoM2	Pr€n0m2	Adr€s\$e2	H	Carrefour	...
NIR 3	01/01/1971	30/11/2019	NoM3	Pr€n0m3	Adr€s\$e3	H	Renault	...
NIR 4	30/04/1996	31/12/2099	NoM4	Pr€n0m4	Adr€s\$e4	F	Sodexo	...
...

...	Adr_Ent	Type_Ctr_Tr	CSP	Deb_AT	Fin_AT	Durée_AT	Motif_AT
...	Lyon	CDI	Cadre	-	-	-	-
...	Villeurbanne	CDI	Cadre	-	-	-	-
...	Nantes	CDD	Salarié	11/02/2019	17/02/2019	7	Maladie
...	Nantes	CDD	Salarié	28/07/2021	30/07/2021	3	Maladie
...	Achères	CDI	Ouvrier	10/01/2017	18/02/2019	771	Accident travail
...	Paris 12ème	CDI	Salarié	20/05/2020	24/05/2020	5	Maladie
...

Nous avons de ce fait notre table d'activité / inactivité provenant des DSN.

Voyons voir comment nous pouvoir fusionner et enrichir nos deux tables...

Conception de la table d'étude / Fusion des tables Santé et DSN

A ce point de l'avancement, nous avons 2 tables à partir de 3 :

- Une table d'Actes en Santé
- Une table d'activité/inactivité provenant des DSN

Nous allons dans un premier temps **fusionner** ces tables afin d'avoir une table permettant l'étude de la **corrélation entre consommations santé et absentéisme**.

Ensuite, nous enrichirons cette fusion avec des **éléments macro pertinents** pour expliquer l'absentéisme : moral des ménages, géographie, moment de l'année. Dans cette partie nous traiterons également les **anomalies** et agrégerons quelques indicateurs pour gagner en efficacité sans perdre d'information.

Critères de jointure entre les tables d'Actes Santé et DSN

La jointure entre nos tables se fera sur l'**individu**.

De plus, la jointure se fera de la manière suivante :

$$\text{Actes Santé} \cap (\text{Activité} + \text{Inactivité DSN})$$

Pour la simple et bonne raison qu'étudier la corrélation entre absentéisme (ou non) et consommation santé perd tout son sens si l'on a des salariés s'arrêtant (ou non) de manière fortuite, inopinée et donc imprédictible.

Ensuite, nous sélectionnerons des consommations d'une antériorité maximale de 2 ans par rapport à chaque élément déclencheur j . Nous sélectionnerons des individus sans anomalie sur le genre, sur les départements (domicile et emploi), et une faible tolérance d'anomalie sur la date de naissance (voir plus bas).

Les arrêts peuvent avoir lieu sur la fenêtre [01/01/2019, 30/04/2022], ce qui impose une fenêtre de consommation santé [01/01/2017, 30/04/2022].

Choix de la maille d'étude

L'objectif de l'étude est de prédire à l'aide des consommations santé dans un premier temps la survenance de l'arrêt de travail, et le cas échéant la gravité de celui-ci.

Pour ce faire, notre variable à prédire sera un indicateur de survenance qui prendra 0 ou 1 selon si l'individu s'est arrêté et un indicateur de gravité qui vaudra la durée d'arrêt en jours.

La maille de la table d'étude sera donc l'arrêt de travail pondéré sur le contrat de

travail, et le contrat de travail de l'individu ne s'étant pas arrêté. Par définition, 1 arrêt de travail est rattaché à 1 seul contrat de travail, puisque les tables DSN ont été retraitées et les chevauchements de contrats de travail ont été éliminés (*cf* : sélection du contrat de travail "cible").

Pondération des lignes

Une règle élémentaire en *Machine Learning* qu'il faut respecter est d'avoir des classes de taille équivalente. Or nous ne pouvons pas attribuer 1 ligne pour chaque arrêt de chaque individu arrêté, et 1 ligne pour 1 contrat de travail sans arrêt de travail. En effet, nous devons prédire la survenance de l'arrêt ; sauf qu'à effectifs équivalents entre cohorte d'individus ayant été inactifs et cohorte d'individus n'ayant pas été arrêtés une seule fois, si le nombre moyen d'arrêts dans le premier groupe est de 5 arrêts sur une période d'activité (comprendre 1 emploi), alors par la loi des Grands Nombres nous aurons environ 5 fois plus de lignes pour la première cohorte que pour la seconde cohorte. Ainsi a été prise l'initiative de pondérer les lignes de la première cohorte par la fraction correspondant à la durée de l'arrêt i sur la durée d'arrêt totale sur cette période d'activité - les lignes de la seconde cohorte auront systématiquement un poids égal à 1. Si on résume, le poids de la ligne j (reliée à l'individu i sur son contrat de travail k) :

$$\delta_j = \begin{cases} \frac{d_{i,j,k}}{d_{i,k}} & \text{si l'individu } i \text{ s'est arrêté sur son contrat } k \\ 1 & \text{s'il ne s'est pas arrêté sur son contrat } k \end{cases}$$

avec $d_{i,j,k}$ la durée de l'arrêt j de l'individu i sur son k -ième contrat de travail et $d_{i,k} = \sum_{j=1}^{n_{arrêts}} d_{i,j,k}$ sa durée totale d'arrêt sur ce contrat de travail.

Ainsi en reprenant l'exemple plus haut, nous aurons un poids de 1 pour chaque contrat de travail présent dans nos tables (avec arrêts ou non).

Agrégation des consommations santé (profil de consommation)

La maille du tableau final étant l'arrêt de travail pour les uns (resp. le contrat de travail pour les autres), nous devons synthétiser toutes les consommations en santé de l'individu en respectant l'antériorité de 2 ans maximum entre la consommation et la survenance de l'arrêt (resp. la fin du contrat de travail). Nous effectuons donc un Tableau Croisé Dynamique du montant pondéré et de la fréquence pondérée par la proximité temporelle à la survenance de l'arrêt (resp. la fin du contrat de travail).

Ce poids fonction de la proximité temporelle sera donc obtenu en traçant les répartitions de la durée entre survenance et la dernière consommation santé, pour chaque poste de consommation santé.

Voici l'allure de ces courbes :

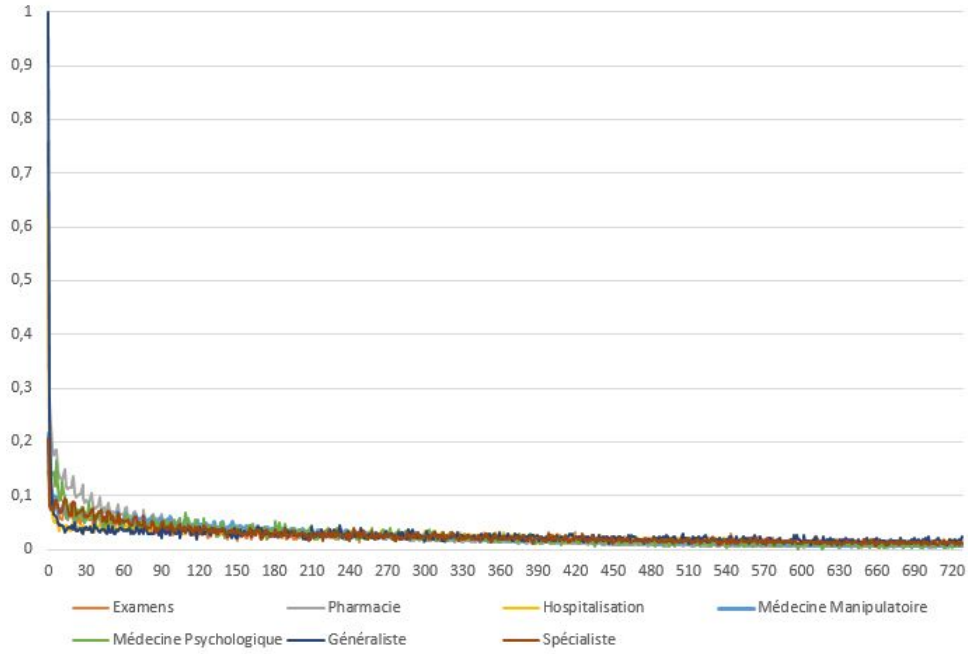


FIGURE 2 – Répartitions de la durée entre survenance et la dernière consommation santé (par poste de consommation), indexées sur la valeur maximale : visite chez le généraliste d'une antériorité de 1 jour

Ces poids étant sortis, nous pouvons définir 14 indicateurs de consommation santé : la fréquence et le montant pondérés par l'antériorité à l'arrêt pour nos 7 postes d'actes Santé.

Notations :

i :	consommation de santé
j :	élément déclencheur (survenance de l'arrêt ou de la fin du contrat si 0 arrêt)
p_i :	poste de consommation de i
t_k :	date de l'évènement k (i ou j)
$a_{i,j}$:	antériorité de i par rapport à j ($a_{i,j} = t_j - t_i$)
$\delta_{p,a}$:	poids d'une consommation de poste p et d'antériorité a (voir courbes)
n_j :	nombre de consommations d'une antériorité maximale de 2 ans par rapport à j (vaut $\sum_{i=1}^{n_j} \mathbb{1}_{a_{i,j} < 730}$)
m_i :	montant de i

D'où pour chaque ligne (1 ligne = 1 élément déclencheur j), la fréquence et le montant pondérés auront cette définition pour le poste de consommation p :

	Fréquence de consommation pour le poste p	Montants de consommation pour le poste p
Évènement déclencheur j	$f_{j,p} = \sum_{i=1}^{n_j} \mathbb{1}_{p_i=p} \delta_{p_i,a_{i,j}}$ (somme des poids des consos)	$m_{j,p} = \sum_{i=1}^{n_j} \mathbb{1}_{p_i=p} \delta_{p_i,a_{i,j}} m_i$ (somme des montants pondérés des consos)

Agrégation des arrêts de travail antérieurs (profil d'inactivité)

Nous allons procéder de la même manière que pour les consommations en santé. En ce qui concerne éléments déclencheurs = {fin du contrat de travail} ou {première survenance d'un arrêt sur le contrat de travail}, l'historique d'inactivité sera logiquement **vide**.

Concentrons-nous maintenant sur les individus s'étant arrêtés au moins 2 fois pendant une période de travail, condition nécessaire et suffisante pour disposer d'un historique d'inactivité. Leur historique d'inactivité comprendra également 2 colonnes par motif d'arrêt (maladie, AT-MP, temps partiel thérapeutique) : la fréquence d'arrêt antérieure et la durée d'arrêt antérieure.

Nous faisons le choix ici aussi de pondérer ces 6 indicateurs par la proximité entre les 2 arrêts : un court délai entre 2 arrêts doit forcément plus peser que 2 arrêts séparés d'une dizaine de mois par exemple.

Le poids attribué à ces deux indicateurs suivra donc la répartition empirique du délai entre deux arrêts, ventilé par les motifs de l'arrêt 1 et 2, ce qui nous donne 9 répartitions :

Voici comment nous pouvons agréger ce profil d'inactivité par rapport à l'arrêt j :

Notations :

$$\left\{ \begin{array}{ll} i : & \text{arrêt de travail antérieur à } j \text{ du même individu X contrat de travail} \\ m_k : & \text{motif de l'arrêt de travail } k \text{ (} i \text{ ou } j \text{)} \\ t_k : & \text{date de survenance de l'arrêt } k \text{ (} i \text{ ou } j \text{)} \\ a_{i,j} : & \text{antériorité de } i \text{ par rapport à } j \text{ (} a_{i,j} = t_j - t_i \text{)} \\ \delta_{m_i,m_j}^{a_{i,j}} : & \text{poids de } i \text{ par rapport à } j \text{ (fonction des motifs et de l'antériorité)} \\ n_j : & \text{nombre d'arrêts } i \text{ antérieurs à } j \end{array} \right.$$

Ainsi, la fréquence et la durée d'arrêt pondérées par rapport à cet arrêt j donnent :

	Fréquence d'arrêt pour le motif m	Durée d'arrêt pour le motif m
Arrêt de travail j de motif m_j	$f_{j,m} = \sum_{i=1}^{n_j} \mathbb{1}_{m_i=m} \delta_{m_i,m_j}^{a_{i,j}}$ (somme des poids des arrêts p/r au motif m_j)	$d_{j,m} = \sum_{i=1}^{n_j} \mathbb{1}_{m_i=m} \delta_{m_i,m_j}^{a_{i,j}} d_i$ (somme des durées des arrêts pondérées p/r au motif m_j)

Traitement des anomalies

Nous avons à notre disposition un certain nombre de variables individuelles dans les deux bases.

L'idée est de garder les données communes aux deux bases et d'écarter les anomalies si elles ne sont pas corrigeables.

Âge à la survenance de l'élément déclencheur j

Pour avoir l'âge à la survenance de j , il faut avoir une date de naissance consolidée. Dans l'immense majorité des cas, les dates de naissance coïncident, mais il subsiste des anomalies de renseignement :

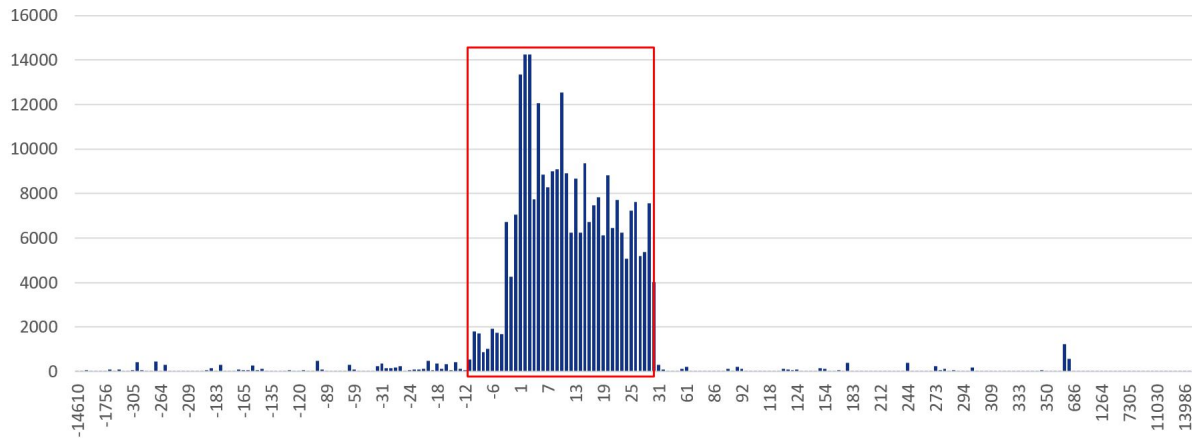


FIGURE 3 – Répartition de la différence (en jours) entre les dates de naissances renseignées dans les deux tables

Nous ferons le choix du côté des anomalies de ne garder que ce qui se trouve dans le rectangle rouge : entre - 11 et + 30 jours d'écart concernant les dates de naissance. En effet, ce rectangle englobe 95% des anomalies et a très peu de chances d'avoir une incidence majeure sur la détermination exacte de l'âge à la survenance (i.e. : dans le pire des cas, on sera à 1 mois près).

La date conservée sera alors la date renseignée dans les DSN.

Genre

Le genre est renseigné dans les 2 tables, mais nous avons des différences parmi elles :

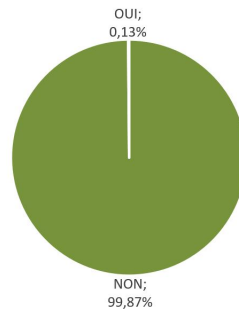


FIGURE 4 – Erreurs sur le genre entre les deux tables

Nous faisons le choix d'écarter les anomalies car elles sont d'un effectif très négligeable (13 pour 10 000) et il est impossible de trancher si une table renseigne un individu masculin et la deuxième le renseigne sous le genre féminin.

Situation familiale

Cette donnée ne figure que dans la table des Actes Santé donc nous la récupérerons telle quelle. Nous allons agréger les 6 modalités initiales pour un individu en 2 modalités qui sont :

Modalité initiale	Modalité finale
Marié	En couple
En concubinage	
Célibataire	Seul
Veuf	
Séparé	
Divorcé	

Catégorie socio-professionnelle

La catégorie socio-professionnelle (CSP) figure dans la table des DSN et nous allons l'agréger en 2 modalités pour alléger la modélisation :

Modalité initiale	Modalité finale
Cadres, professions intellectuelles supérieures	Cadre
Agriculteurs	Non cadre
Artisans, commerçants et chefs d'entreprise	
Professions intermédiaires	
Employés	
Ouvriers	
Autres	

Région

La variable géographique de nos salariés étant le Code Postal, on risque de trop complexifier le modèle, il est donc nécessaire de regrouper nos salariés par région, mais cela fait encore beaucoup de modalités. Il convient donc de fusionner nos régions administratives en régions ayant une démographie équivalente et à profils + géographie similaires. Le découpage proposé est le suivant :

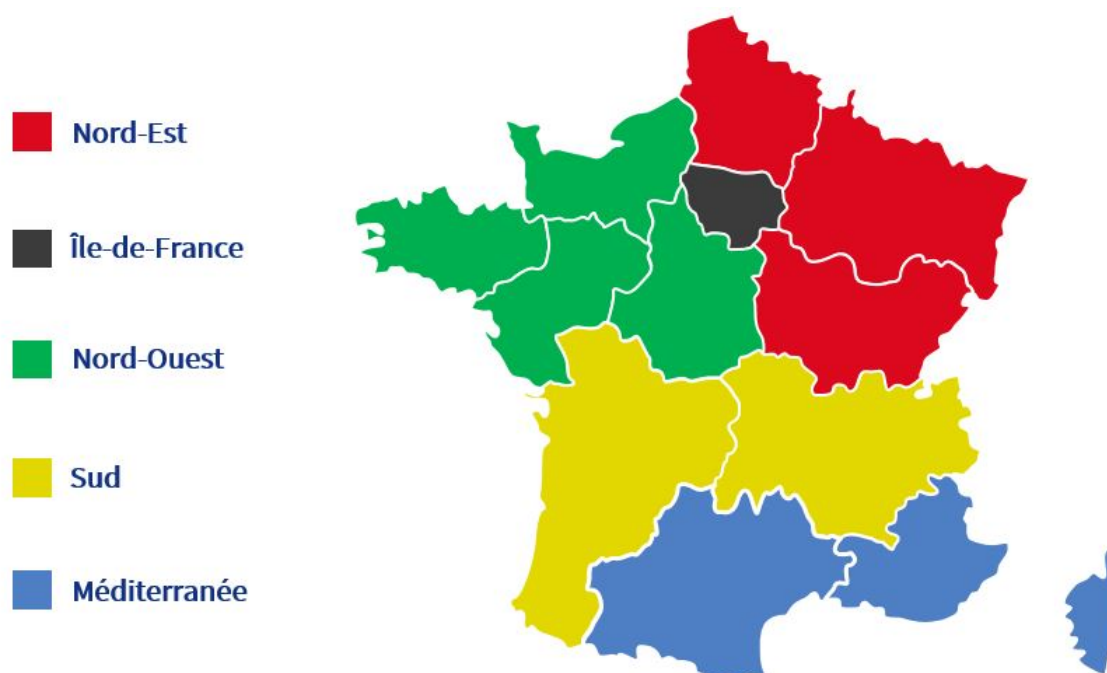


FIGURE 5 – Agrégation des régions des salariés

Type de contrat de travail

Le type de contrat de travail figure dans les données des DSN, avec cette répartition :

Libellé	Proportion
Contrat de travail à durée indéterminée de droit privé	91,97%
Contrat de travail à durée déterminée de droit privé	7,74%
Mandat social	0,11%
Contrat de soutien et d'aide par le travail	0,06%
Autre nature de contrat, convention, mandat	0,04%
Contrat de travail à durée indéterminée de chantier ou d'opération	0,02%
Contrat d'engagement maritime à durée indéterminée	0,01%

Nous ne garderons que les deux premiers libellés, qui forment à eux deux 99,7% de notre base, et nous les renommerons « CDI » et « CDD ».

Secteur d'emploi

Le secteur d'emploi est renseigné dans les données des DSN, nous pouvons récupérer le secteur d'emploi de chaque salarié aisément.

Taille de l'entreprise

La taille de l'entreprise est une donnée qu'il faut calculer à partir des données des DSN.

Nous n'avons que la dernière actualisation (avril 2022) pour les DSN donc nous n'aurons pas une vision exacte de la population de l'entreprise au mois de la survenance de l'évènement déclencheur, mais la vision au 30 avril 2022 (ce qui est un indicateur déjà pertinent).

La méthode de calcul est donc de sommer le nombre de salariés distincts d'un même SIREN (entreprise au sens large). Ce nombre est ensuite adjoint à chacun de nos salariés avec la clé de jointure « SIREN ».

Création d'indicateurs macro

Distance domicile - travail

La mise en place brutale du télétravail avec la crise Covid a fait ressortir l'importance du bien-être, avec entre autres la flexibilité de s'organiser, de pouvoir choisir ses jours de « présentiel ».

Le télétravail séduit donc naturellement beaucoup de salariés habitant loin de leur emploi, et il serait intéressant de voir s'il y a une corrélation entre la distance domicile - travail et la survenance d'un arrêt de travail ou l'arrêt d'un contrat de travail.

Pour ce faire, nous avons transformé tous les codes postaux des salariés et des entreprises en coordonnées GPS, puis avons appliqué la formule d'Haversine [10] pour obtenir la distance entre deux points du globe :

$$d(x_1, x_2) = 12742 \arcsin \sqrt{\sin\left(\frac{lat_2 - lat_1}{2}\right)^2 + \cos(lat_1) \cos(lat_2) \sin\left(\frac{long_2 - long_1}{2}\right)^2}$$

Cette distance domicile - travail est ensuite adjointe à la table en tant que caractéristique individuelle. Nous exclurons par principe les distances trop grandes : habitants en outre-mer et salariés en métropole (ou inversement). Ce sont des outliers qui fausseront le modèle plus qu'ils n'apportent de l'information ; de plus ils ne représentent que 0,012% de la table.

Nous manquons cependant de précision car dans des grandes villes autres que Paris, Lyon et Marseille qui ont des Codes Postaux par arrondissements, cette donnée n'est pas très représentative de la distance, il peut nécessiter de rouler pendant 20 minutes alors que l'on reste dans le même arrondissement (et donc même code postal, distance = 0).

Voici comment y remédier...

Zone dense

Nous allons sélectionner les 15 villes les plus embouteillées de France à l'aide du site TomTom [11] :

N°	Ville	Taux de congestion en 2021
1	Paris	36%
2	Marseille	35%
3	Toulon	33%
4	Bordeaux	32%
5	Lyon	29%
6	Nice	28%
7	Montpellier	27%
8	Grenoble	27%
9	Strasbourg	26%
10	Nantes	25%
11	Brest	24%
12	Rennes	24%
13	Toulouse	23%
14	Le Havre	23%
15	Lille	22%

Tous les codes postaux associés à ces villes seront donc estampillés d'une indicatrice de zone dense, signifiant que la distance domicile - travail doit être regardée d'un autre œil.

Variables propres à la date de survenance de l'élément déclencheur j

Nous allons dans cette partie nous pencher sur quels facteurs temporels peuvent expliquer la survenance d'un élément déclencheur j .

Un arrêt qui survient un lundi ou un vendredi peut être un moyen de prolonger le week-end, un arrêt qui survient autour d'un jour férié ou de vacances scolaires idem. Ces indicateurs doivent être étudiés et voir s'il y a réellement une corrélation. La saison peut être également étudiée.

Distance au week-end

Pour trouver la distance au week-end, le moyen est de créer à la main une table avec chaque date du 01/01/2017 au 30/04/2022 (fenêtre de notre étude) et de renseigner la donnée.

Cette manipulation est assez facile puisqu'il suffit de répéter un vecteur de 7 éléments (0, 0, 1, 2, 3, 2, 1) en le calibrant sur le premier week-end de la fenêtre.

Distance à un jour férié ou à des vacances scolaires

La distance à un jour férié ou aux vacances scolaires elle, n'est pas particulièrement dure à calculer mais il faut faire attention à avoir toutes les zones géographiques : les

académies type zone A, B, C qui ont des calendriers scolaires différents.

La distance s'exprime alors ainsi (jv = jour vaqué) :

$$d_{jv}(t_j) = \min(t_{jv_suivant} - t_j, t_j - t_{jv_précédent})$$

Saison de l'année

La saison de l'année correspond à la saison du jour de l'évènement déclencheur. Cette variable prend logiquement 4 modalités : Été, Automne, Hiver, Printemps.

Moral des ménages

L'indicateur du moral des ménages est fourni par l'INSEE [12]. Il synthétise l'opinion des Français sur la situation économique. Plus il est élevé, plus la confiance des ménages dans la situation économique est favorable.

Il est calculé à l'aide d'une analyse factorielle (qui résume l'évolution de plusieurs variables dont les mouvements sont corrélés).

Cet indicateur est donc le fruit de l'analyse factorielle de 8 soldes d'opinions (% réponses positives - % réponses négatives) :

- Niveau de vie passé
- Niveau de vie futur
- Situation financière passée
- Situation financière future
- Chômage
- Opportunité de faire des achats importants
- Capacité d'épargne actuelle
- Capacité d'épargne future

Sont ainsi interprétés les évolutions et les écarts à la moyenne de long-terme : d'où des valeurs au-delà de 100 lorsque la situation est favorable.

Près de 2 000 ménages sont interrogés chaque mois.

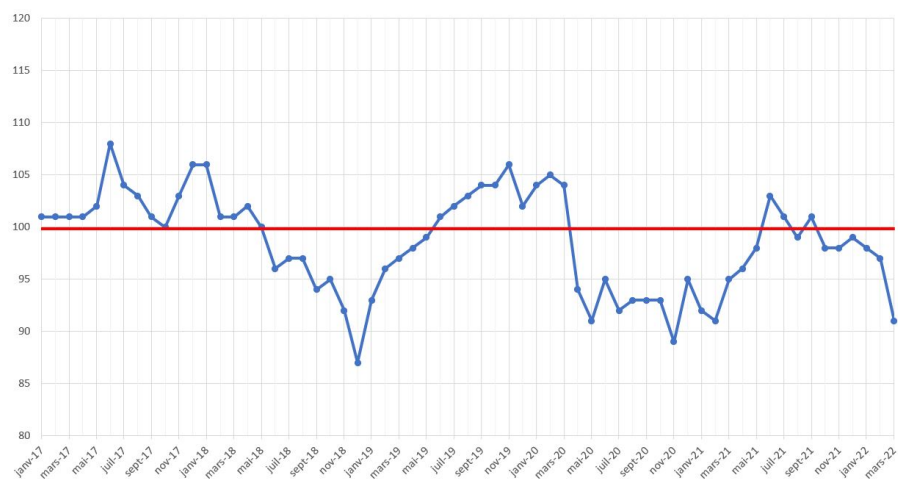


FIGURE 6 – Moral des ménages français ([source : INSEE](#))

Cette valeur mensuelle est donc adjointe au tableau en prenant pour référence le mois de survenance de l'évènement déclencheur j .

Statistiques descriptives

Cette partie a pour but de présenter quelques statistiques descriptives.
La table que nous avons obtenue contient 793 897 lignes, dont :

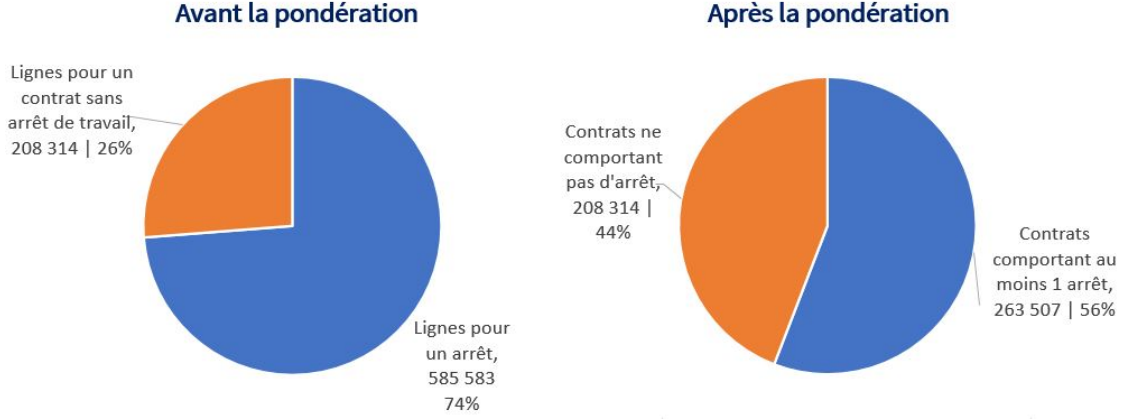


FIGURE 7 – Répartition des lignes brutes selon la survenance ou non d'arrêt de travail (gauche), et répartition des contrats selon survenance d'arrêt de travail ou non (droite)

Le second graphe est donc le premier graphe avec les pondérations permettant d'accorder autant d'importance à un salarié qui s'est arrêté qu'à un salarié ne s'étant pas arrêté une seule fois (durée d'arrêt i / durée d'arrêt totale). Cette agrégation a donc permis de passer d'une répartition 74-26 à une répartition 56-44 sans perdre d'information.

De plus, on retrouve taux d'absentéisme réaliste puisqu'on est sur 3 ans et 4 mois d'observation :

Soit A_i l'évènement être absent au moins une fois lors de l'année i . La probabilité d'être absent au moins une fois sur la période 01/2019-04/2022 vaudra donc :

$$\begin{aligned}
 \mathbb{P}(A_{2019-04/2022}) &= 1 - \mathbb{P}(\overline{A_{2019-04/2022}}) \\
 &= 1 - \mathbb{P}(\overline{A_{2019}} \cap \overline{A_{2020}} \cap \overline{A_{2021}} \cap \overline{A_{04/2022}}) \\
 &= 1 - \mathbb{P}(\overline{A_{2019}}) * \mathbb{P}(\overline{A_{2020}}) * \mathbb{P}(\overline{A_{2021}}) * \mathbb{P}(\overline{A_{04/2022}}) \\
 &= 1 - \left((1 - \mathbb{P}(A_{2019})) * (1 - \mathbb{P}(A_{2020})) * (1 - \mathbb{P}(A_{2021})) * (1 - \mathbb{P}(A_{04/2022})) \right) \\
 &= 1 - \left((1 - 0,32) * (1 - 0,35) * (1 - 0,34) * (1 - 0,11) \right) \\
 &= 1 - (0,68 * 0,65 * 0,66 * 0,89) \\
 &= 0,74
 \end{aligned}$$

NB : Nous faisons le proxy suivant (on prend la moyenne de l'absentéisme des 3

années précédentes et on applique le *prorata*) :

$$\mathbb{P}(A_{04/2022}) = \frac{4}{12} \hat{\mathbb{P}}(A_{2022}) = \frac{4}{12} \frac{\mathbb{P}(A_{2019}) + \mathbb{P}(A_{2020}) + \mathbb{P}(A_{2021})}{3} = 0,11$$

On retombe donc exactement sur notre répartition 74-26.

Nous pouvons maintenant nous intéresser au nombre d'arrêts de travail recensés par contrat de travail individualisé et voir comment la catégorie socio-professionnelle, la tranche d'âge ou le genre influe dessus :

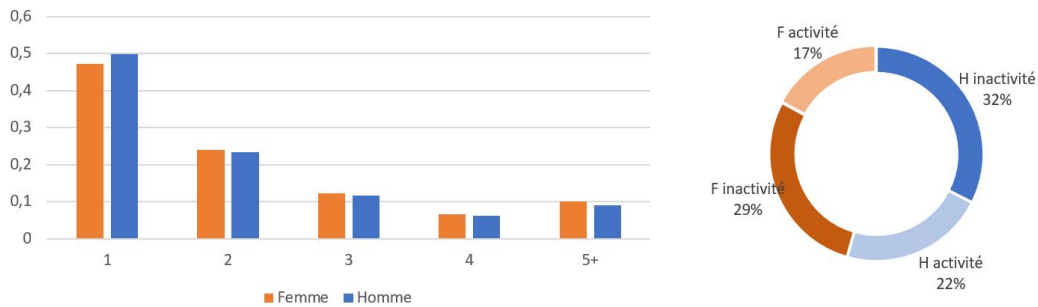


FIGURE 8 – Répartition du nombre d'arrêt par genre (gauche) et démographie (droite)

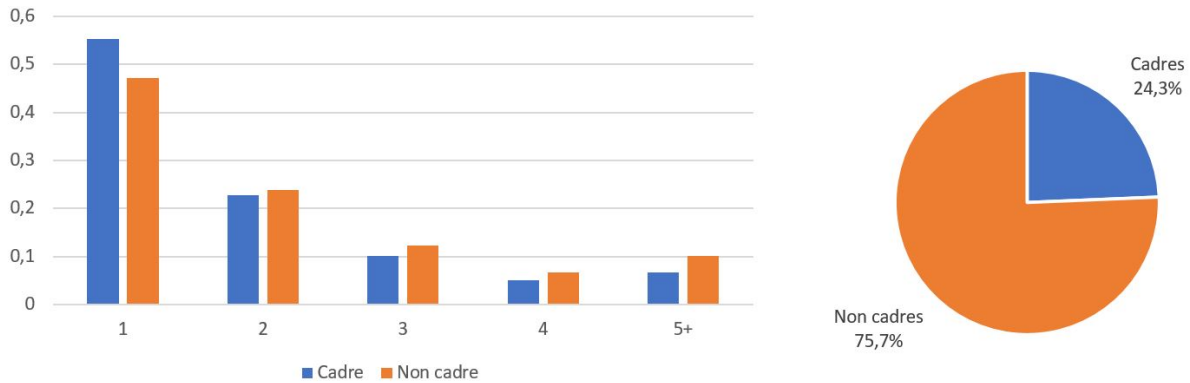


FIGURE 9 – Répartition du nombre d'arrêt par CSP (gauche) et démographie (droite)

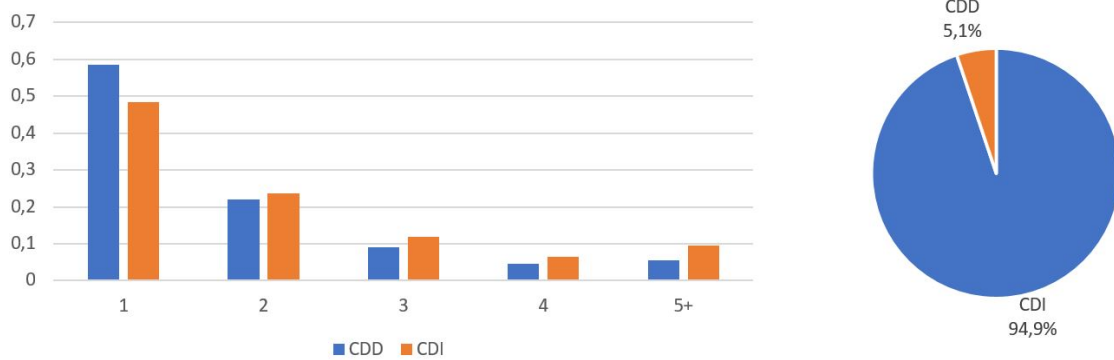


FIGURE 10 – Répartition du nombre d'arrêt par contrat (gauche) et démographie (droite)

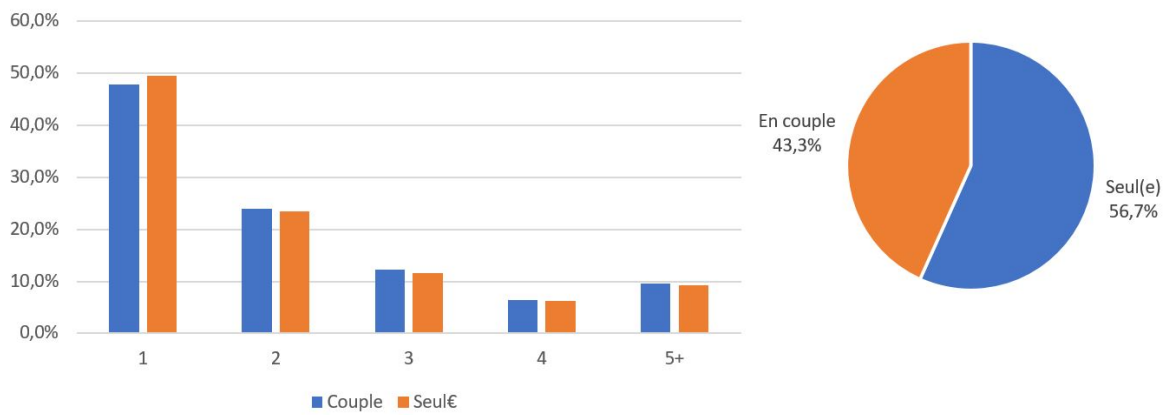


FIGURE 11 – Répartition du nombre d'arrêt par situation familiale (gauche) et démographie (droite)

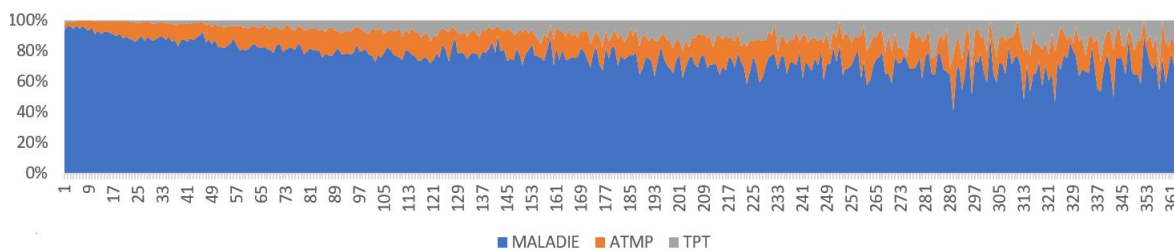


FIGURE 12 – Proportion des motifs d'arrêt selon la durée de l'arrêt

Modélisation du risque incapacité

Dans le cadre de notre étude, nous allons opter pour 2 modèles consécutifs. Le premier doit servir à prévenir la survenance (qui est une variable binaire). Il ne doit donc que reposer sur les consommations santé, les variables individuelles (type âge, genre, situation familiale), les variables propres à leur emploi (taille entreprise, distance domicile-emploi, zone dense ou non, type de contrat, CSP, secteur d'activité) et des variables macro de l'élément déclencheur j (moral des ménages, distance aux vacances/jours fériés, au week-end, saison). Le second modèle servira à prédire la durée de l'arrêt de travail. Celui-ci peut donc prendre en input les mêmes variables, ainsi que l'historique d'inactivité des individus arrêtés.

Prédiction de la survenance d'un arrêt de travail

Pour prédire la survenance, nous avons opté pour plusieurs modèles. Tout d'abord, vu le caractère binaire de la sortie, nous avons choisi de faire une régression logistique avec toutes nos variables. En règle générale, la modélisation de l'incapacité de travail se fait par les moyens de **modèles linéaires généralisés** (GLM), et notamment par la **régression logistique**.

Dans un premier temps nous introduirons les GLM puis la régression logistique (qui est un GLM en particulier), puis dans une seconde section nous verrons une application de la régression logistique sur nos données.

Généralités sur les modèles linéaires

Lorsque l'on étudie une variable Y à l'aide de variables X^1, \dots, X^p , on peut aboutir à un modèle linéaire multivarié ; si l'on a n observations [13] :

$$\forall i \in \{1, \dots, n\}, \quad Y_i = \beta_0 + \beta_1 X_i^1 + \dots + \beta_p X_i^p + \epsilon_i$$

Ce qui revient à écrire à l'aide de matrices :

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

Dans notre cas, Y représenterait le vecteur durée de l'arrêt de travail, $\beta = (\beta_0 \beta_1 \dots \beta_p)$ nos prédicteurs linéaires, $X = (\mathbb{1} X^1 \dots X^p)^T$ nos variables explicatives (consommations santé, géographie, CSP, distance domicile-entreprise, etc.).

Cette modélisation est utilisée pour :

- **La description** : ce modèle décrit la relation $Y = f(X)$
- **La contributions** : ce modèle fournit les pondérations de chaque variable
- **prédiction** : ce modèle est très facile à mettre en place pour prédire de nouvelles valeurs

Cependant, le modèle linéaire dit « classique » manque de flexibilité et est surtout un tremplin pour l'utilisation beaucoup plus commode des modèles linéaires « généralisés ».

Les modèles linéaires généralisés (GLM) permettent d'étudier la corrélation entre une variable réponse Y et plusieurs variables explicatives X^1, \dots, X^p . Ils englobent :

- le modèle linéaire « classique »
- la régression logistique
- le modèle log-linéaire
- la régression de Poisson

Un GLM est constitué de 3 termes :

- **Terme aléatoire** : la densité de la loi Y_i (échantillon de la variable réponse Y) appartient à la famille exponentielle et est de la forme :

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

avec a , b des fonctions spécifiques à chaque famille exponentielle et c une fonction quelconque.

- **Terme déterministe** : précise quels sont les prédicteurs linéaires

$$\eta(x_i) = \beta_0 + \beta_1 x_i^1 + \dots + \beta_p x_i^p$$

- **Fonction de lien** : fonction inversible g telle que

$$g(\mathbb{E}[Y_i]) = \eta(x_i) = \beta_0 + \beta_1 x_i^1 + \dots + \beta_p x_i^p$$

Voici des exemples de fonctions de lien :

Distribution	Utilité	Fonction de lien $g(y)$	$a(x)$	$b(x) (g = (b')^{-1})$
$N(\mu, \sigma^2)$	Valeurs normales	y	σ^2	$\frac{x^2}{2}$
$P(\lambda)$	Comptage	$\log(y)$	1	e^x
$Bin(n, p)$	Pourcentage	$logit(y) = \log(\frac{y}{n-y})$	n	$\log(n + e^x)$
$\Gamma(\mu, \nu)$	Durée	$-\frac{1}{y}$	$\frac{1}{v}$	$-\log(-x)$

Théorie de la régression logistique

Si notre variable réponse Y est binaire, l'objectif de la régression logistique est d'estimer les probabilités [14] :

$$\begin{cases} p_1(x) = \mathbb{P}(Y = 1 \mid (X_1, \dots, X_n) = x) \\ p_0(x) = \mathbb{P}(Y = 0 \mid (X_1, \dots, X_n) = x) = 1 - p_1(x) \end{cases}$$

La régression logistique repose sur l'hypothèse suivante :

$$\ln \left(\frac{p_1(x)}{p_0(x)} \right) = a_0 + a_1 x_1 + \dots + a_n x_n$$

Comme $p_0 = 1 - p_1$, l'hypothèse devient :

$$\ln \left(\frac{p_1(x)}{1 - p_1(x)} \right) = logit(p_1(x)) = a_0 + a_1 x_1 + \dots + a_n x_n$$

Cette équation, que l'on appelle **log-odd-ratio** donne, par inversion de la fonction *logit* :

$$\begin{cases} p_1(x) = \frac{\exp(a_0 + a_1 x_1 + \dots + a_n x_n)}{1 + \exp(a_0 + a_1 x_1 + \dots + a_n x_n)} \\ p_0(x) = \frac{1}{1 + \exp(a_0 + a_1 x_1 + \dots + a_n x_n)} \end{cases}$$

La section suivante s'attachera à présenter les résultats de notre modélisation par **régression logistique**.

Modélisation de la survenance de l'arrêt de travail par régression logistique

Nous allons fit ce modèle de régression logistique sur 80% des valeurs prises au hasard (échantillon *train*) puis le backtester sur les 20% de valeurs restantes (échantillon *test*). Cette méthode d'entraînement de modèles appelée « train-test split » [15] est très efficace quand on a un très grand nombre de données. Elle est simple d'utilisation et permet de réduire les risques de sur-apprentissage.

L'output de cette modélisation sur les données *train* est trouvable en Annexe 1.

Cette sortie n'est pas acceptable, déjà à cause de la très forte $Pr(> |z|)$ de l'Intercept, mais aussi car aucune modalité de la variable catégorielle *SECTEUR* n'est pertinente au regard de cette statistique.

On va donc en profiter pour écarter toutes les variables non-significatives, à commencer par le *SECTEUR*, l'historique des ATMP et des Temps Partiels Thérapeutiques passés (on ne garde que l'historique des arrêts maladie passés), la consommation en pharmacie, médecine spécialiste et examens.

Nous obtenons un output tout à fait crédible en **Annexe 2**.

Ce modèle que l'on va backtester sur les 20% de données *test* nous sort des valeurs dont l'écrasante majorité se situe entre 0 et 1 (une petite partie sont négatives, une autre petite partie sont > 1).

On applique alors une **classification bayésienne naïve**, calibrée dans un premier temps à 0,5 afin de séparer les valeurs continues en binaire 0-1 :

$$\begin{cases} 1 & \text{si pred} \geq 1/2 \\ 0 & \text{si pred} < 1/2 \end{cases}$$

Voici la matrice de confusion de cette prédiction :

		Réel	
		1	0
Prédit	1	101 422 63,9%	11 308 7,1%
	0	15 831 10%	30 219 19%

Le tableau est de cette forme :

Vrai Positif	Faux Positif
Faux Négatif	Vrai Négatif

On a donc une sensibilité de $S_e = \frac{VP}{VP+FN} = 86,5\%$ (probabilité de bien prédire la survenance 1 d'un arrêt de travail), une spécificité de $S_p = \frac{VN}{VN+FP} = 72,8\%$ (probabilité de bien prédire l'absence 0 d'arrêt de travail) et un R^2 de 0,854, ce qui est complètement acceptable d'autant plus que nous aurons besoin pour la suite de modéliser la gravité, qui est sous-jacente à une survenance 1.

Nous avons encore un levier pour améliorer nos précisions : la valeur de 1/2 étant arbitraire, on peut la faire naviguer dans $[0,1]$ pour voir en quel point elle donne un R^2 maximal.

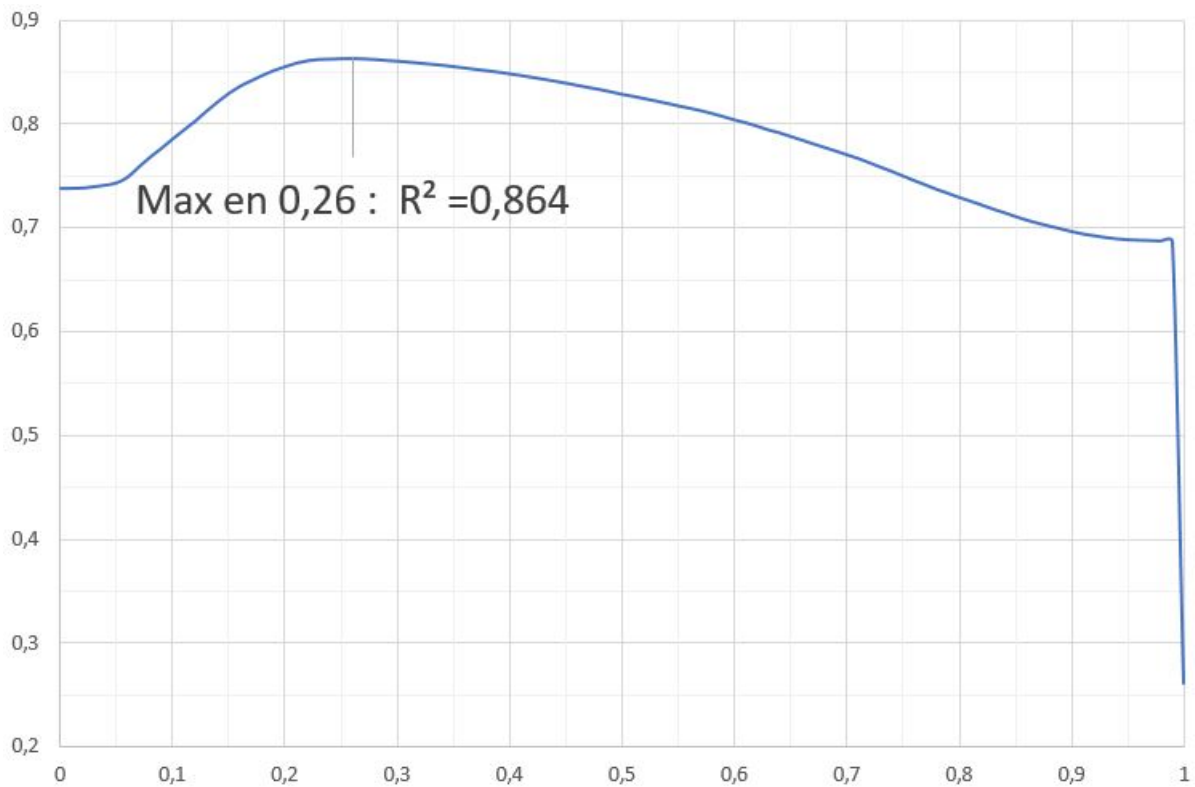


FIGURE 13 – Coefficient de corrélation du Bayésien Naïf en fonction du seuil d'attribution 0-1

Sans surprise, la valeur maximisant le R^2 est atteinte en 0,26, qui correspond exactement à la frontière dans la répartition de nos classes « **Non-survenus** / **Survenus** ». Voici les éléments précédents avec cette fois un seuil à 0,26 au lieu de 0,5 :

		Réal	
		1	0
Prédit	1	114 506 72,1%	18 915 11,9%
	0	2 747 1,7%	22 612 14,2%

Et la **sensibilité**, **spécificité**, et R^2 qui sont associés :

Sensibilité	97,7%
Spécificité	54,5%
R^2	0,864

Même si nous perdons en spécificité, nous avons une modélisation très précise quant aux vrais positifs et qui est également prudente puisque nous nous attendons à $\sim 19\,000$ arrêts qui ne se sont pas produits au regard des $\sim 2\,750$ arrêts qui ont échappé à

l'acuité du modèle.

Cette modélisation bien que satisfaisante, mérite d'être comparée avec des méthodes de *Machine Learning*, réputées pour leurs performances.

Random Forest

Avant de parler de l'algorithme des forêts aléatoires [14] (Random Forest en anglais), nous devons introduire les arbres de classification et régression (CART en anglais).

Un arbre est un algorithme d'apprentissage supervisé. Son but est de classer la population en 2 catégories selon un test, par exemple : « L'individu est-il Cadre ? ». L'individu est ensuite classé dans l'un des deux groupes appelés « noeuds ».

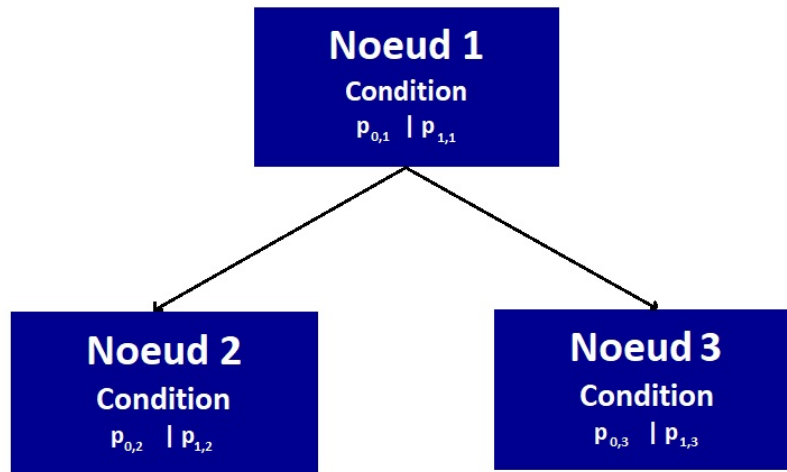


FIGURE 14 – Exemple d'arbre de décision

La condition de séparation est celle qui maximisera l'hétérogénéité des deux groupes :

$$H(k) = p_{0,k} * (1 - p_{0,k}) + p_{1,k} * (1 - p_{1,k})$$

avec p_i, k la proportion de $i \in \{0, 1\}$ pour la condition k .

Cet algorithme peut boucler jusqu'à obtenir ce qu'on appelle « arbre maximal » : chaque individu est classé exactement au bon endroit. Cependant, l'arbre maximal est synonyme de **sur-apprentissage**, ce qui signifie que l'algorithme ne sera pas adapté à de nouvelles valeurs.

De ce fait, il faut élaguer l'arbre pour obtenir l'arbre optimal (en faisant jouer le paramètre de complexité cp).

L'avantage de ce modèle est qu'il n'exige pas l'indépendance des variables explicatives,

donc aucun retraitement n'est requis. Cependant, cet algorithme est plutôt instable, un changement léger sur l'arbre risque de radicalement changer l'algorithme ; ce pour quoi nous allons utiliser l'algorithme Random Forest.

Cet algorithme, introduit en 2001 par Breiman est un agrégat d'arbres de décision, d'où la métaphore. Le principe est le suivant :

1. Créer B paquets d'apprentissage avec à chaque fois des individus différents (bootstrap) **et** des variables explicatives tirées au sort (**principe du Bagging**)
2. Exécuter l'algorithme CART sur chaque paquet
3. Récupérer la prédiction finale de chaque individu à prédire en fonction du vote majoritaire sur les B algorithmes

En clair, Random Forest effectue un apprentissage sur plusieurs arbres de décision entraînés sur des sous-ensembles de données légèrement différents. Il est de ce fait nettement plus stable que l'algorithme CART.

On peut aussi définir l'erreur **Out-of-bag** (OOB) qui est l'erreur de prédiction moyenne des individus lorsqu'ils ne se retrouvent pas sélectionnés dans le paquet d'apprentissage.

Ce modèle Random Forest nous fournit de très bons résultats, encore meilleurs qu'avec la régression logistique :

		Réel	
		1	0
Prédit	1	115 540 72,8%	14 351 9%
	0	1 713 1,1%	27 176 17,1%

Voici les éléments de cette prédiction :

Sensibilité	98,5%
Spécificité	65,4%
R^2	0,899

Ce modèle est donc assurément robuste.

XGBoost

Méthodes de Boosting

Le **Boosting** [14] repose sur le même principe que le « Bagging », processus que l'on a vu dans le Random Forest où on sélectionne certains individus de manière bootstrapée et certaines colonnes. A l'inverse de ce dernier, les algorithmes du Boosting sont

dépendants les uns des autres puisque chacun est entraîné sur son prédécesseur en affectant un poids plus important aux données mal prédites : cela permettra de corriger ces erreurs.

Contrairement au Random Forest, les arbres de décision sont créés en série.

Gradient Boosting

Qui dit **Gradient Boosting** dit **descente de gradient** : il s'agit d'un algorithme d'optimisation différentiable, qui permet de trouver le minimum de n'importe quelle fonction convexe en convergeant progressivement vers celui-ci. Cela est intéressant dans notre cas car la fonction de **coût** est justement convexe.

Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction différentiable, $\eta \geq 0$ un taux d'apprentissage et a_0 un point initial. Le gradient d'une fonction f est défini comme :

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

L'algorithme de descente de gradient est ainsi défini :

1. Calculer $\nabla f(a_t)$
2. Calculer $a_{t+1} = a_t - \eta \nabla f(a_t)$
3. Répéter 2 et 3 jusqu'à la condition de sortie ($\|\nabla f(a_t)\| = 0$ ou nombre d'itérations max atteint)

Le choix d' a_0 est important puisqu'on pourrait aboutir à un **minimum local** différent du minimum de la fonction f .

Le Gradient Boosting va trouver une combinaison linéaire d'arbres optimale en construisant une nouvelle base corrélée avec le gradient de la fonction de coût. Cette technique repose donc sur un mélange de Boosting et de descente de gradient.

Principe du XGBoost

XGBoost, pour eXtreme Gradient Boosting représente une généralisation du Gradient Boosting. Il a l'avantage de minimiser le biais tout en maîtrisant la variance.

De plus, **XGBoost** traite les données en plusieurs blocs compressés, et ce en parallèle ce qui a pour but de drastiquement réduire les temps de calcul.

Enfin, l'**XGBoost** permet de modifier un certain nombre de paramètres afin de notamment limiter le risque de sur-apprentissage ou de choisir la fonction de coût. Il peut être utilisé pour de la classification comme pour de la régression.

Dans la prochaine section seront présentés les résultats du **XGBoost**...

		Réel	
		1	0
Prédit	1	110 649 69,7%	9 674 6,1%
	0	6 604 4,2%	31 853 20,1%

Voici les éléments de cette prédiction :

Sensibilité	94,4%
Spécificité	76,7%
R ²	0,897

Les modèles Random Forest et XGBoost sont deux très bons modèles, on le voit encore ici.

Nous devons peser le pour et le contre de nos 3 modèles : la **régression logistique** nous fournit des résultats très acceptables mais avec une assez grosse sur-estimation de la survenance, et les modèles de **Machine Learning** sont performants surtout en terme de spécificité, mais ont un caractère **boîte noire**. Une idée pourrait être de se servir de cette forte capacité à discerner les vrais négatifs pour renforcer notre régression logistique et améliorer nos résultats ; ce que nous allons voir dans la section suivante...

Amélioration de la modélisation en combinant les méthodes

L'idée alors pourrait être de *fit* une régression logistique comme initialement, sélectionner les positifs prédits (97% de l'effectif positif + 45% de l'effectif négatif!), puis appliquer les modèles Random Forest (resp. XGBoost) ajustés sur la base d'entraînement complète car ceux-ci possèdent d'excellentes spécificité/sensibilité : cela permettrait alors de discerner avec une très grande acuité les vrais positifs (positifs après régression logistique **et** algorithme de Machine Learning) des faux positifs de la régression logistique (positifs après régression logistique et négatifs après algorithme de Machine Learning → ces derniers seront donc *topés* négatifs).

Nous sélectionnons donc les positifs provenant de la régression logistique sur la table de test (133 421 observations : 114 506 Vrais Positifs + 18 915 Faux Positifs).

Régression Logistique puis Random Forest

Cette combinaison de méthode nous donne cette nouvelle matrice de confusion :

		Réel	
		1	0
Prédit	1	114 229 71,9%	2 597 1,6%
	0	3 024 1,9%	38 930 24,5%

Voici les éléments de cette prédiction :

Sensibilité	97,4%
Spécificité	93,7%
R ²	0,965

Les résultats sont extrêmement bons, il nous reste maintenant à voir si le XGBoost après la régression logistique fera mieux.

Régression Logistique puis XGBoost

Cette combinaison de méthode nous donne cette nouvelle matrice de confusion :

		Réel	
		1	0
Prédit	1	109 409 68,9%	8 035 5,1%
	0	7 844 4,9%	33 492 21,1%

Voici les éléments de cette prédiction :

Sensibilité	93,3%
Spécificité	80,7%
R ²	0,90

Les résultats n'ont que très peu évolué grâce au modèle XGBoost (ils restent très bons).

Choix de la modélisation de la survenance de l'arrêt de travail

La modélisation de la survenance de l'arrêt de travail est la première pierre de notre travail, mais sa précision est décisive pour le futur i.e. la prédiction de la durée des arrêts survenus.

Nous avons au départ eu de bons résultats avec une régression logistique, que nous

avons ajustée dans un second temps pour maximiser le R^2 . Nous avons confronté cette méthode à 2 algorithmes de Machine Learning très puissants.

Ils ont en effet fourni de meilleurs résultats mais ont un certain caractère **boîte noire**. De ce fait, nous avons combiné la régression logistique (qui nous fournissait une proportion non-négligeable de Faux Positifs) avec chacun des algorithmes de Machine Learning pour améliorer nos résultats en ce qui concerne la proportion de Faux Positifs.

Voici ce que nous pouvons en tirer :

Modèle	R^2	Sensibilité	Spécificité
Régression Logistique #1	0,829	86,5%	72,8%
Régression Logistique #2	0,864	97,7%	54,5%
Random Forest	0,899	98,5%	65,4%
XGBoost	0,897	94,4%	76,7%
Régression Logistique #2 puis Random Forest	0,965	97,4%	93,7%
Régression Logistique #2 puis XGBoost	0,900	93,3%	80,7%

Le modèle retenu sera donc la **Régression Logistique #2 puis Random Forest**, au vu des performances qui sont soit maximales soit proches de l'être.

Validation de ce modèle

Nous allons tester ce modèle combiné sur plusieurs échantillons en appliquant la méthode du **Bootstrap**. Nous allons demander une table avec 50% d'arrêts survenus (1) et 50% d'arrêts non-survenus (0) piochés au hasard dans notre table d'étude.

Nous avons donc **53 846 lignes positives** et **53 846 lignes négatives**. Le modèle nous donnera la matrice de confusion suivante :

		Réel	
		1	0
Prédit	1	53 075 49,28%	3 356 3,12%
	0	771 0,72%	50 490 46,88%

Ce qui nous donne alors :

Sensibilité	94%
Spécificité	98,5%
R^2	0,962

Ce modèle ne sur-apprend donc pas et **est retenu pour modéliser la survenance d'un arrêt de travail**. Dans la partie suivante seront présentées les contributions du modèle.

Interprétation du modèle

Odd-ratios de la régression logistique

Variable	Modalité	Odd-ratio
Conso hospi.	[0,46436.9]	0
Conso généraliste	[0,9125]	0
Conso méd. manip.	[0,4330.4]	0
Conso méd. psycho	[0,13580.7]	0
Âge	[15,93]	0
Genre	F	réf
	M	-0.03
Situation familiale	Couple	réf
	Seul(e)	-0.04
CSP	Cadre	réf
	Non cadre	0.09
Contrat	CDD	réf
	CDI	0.27
Taille entreprise	ETI	réf
	GE	0.07
	TPE/PME	-0.05
Moral	[88,106]	0.01
Distance jour vaqué	[0,27]	0.01
Distance au we	0	0
	1	0
	2	0
	3	0
Saison	Automne	réf
	Été	-0.02
	Hiver	-0.01
	Printemps	-0.16
Historique A maladie	[0,219]	0.01
Région	IDF	réf
	Méditerranée	0.01
	Nord-Est	0.03
	Nord-Ouest	0.02
	Sud	0.02

Rappel : l'odd-ratio de la variable **Contrat** présent dans la tableau ci-avant se lit de la manière suivante :

$$OR = \frac{P(\text{Survenance} = 1 | \text{Contrat} = \text{CDI})}{P(\text{Survenance} = 0 | \text{Contrat} = \text{CDI})} \bigg/ \frac{P(\text{Survenance} = 1 | \text{Contrat} = \text{CDD})}{P(\text{Survenance} = 0 | \text{Contrat} = \text{CDD})} = 1 + 0.27 = 1.27$$

Il faut donc l'interpréter de la manière suivante : « si l'on compare le rapport de risque entre tomber en arrêt de travail et ne pas tomber en arrêt de travail pour un CDI avec le même rapport de risque pour un CDD, celui des CDI est 1,27x plus élevé ». La sinistralité est donc plus élevée au sein de la cohorte **CDI**.

Ainsi d'après la régression logistique, les modalités les plus sinistrées en terme de survenance d'un arrêt de travail serait le fait d'être en **CDI** (+27%), **Non cadre** (+9%), dans une **Grande Entreprise** (+7%)...

Remarque : la variable **Saison** propose un résultat biaisé... Si on lisait le tableau naïvement, on pourrait croire que le **Printemps** est une saison beaucoup moins risquée (-16%) au niveau des arrêts de travail. L'explication à cela est la fin de notre intervalle temporel d'étude (30/04/2022) ce qui a impliqué de renseigner comme date de survenance de l'élément perturbateur j le 30/04/2022 pour un certain nombre d'individus sans arrêt de travail.

Features Importance du Random Forest (Gini Mean Decrease)

Le modèle Random Forest s'avère être puissant mais ne fournit que très peu d'informations sur les contributions de chaque variable, puisque les variables sont utilisées sous forme de tests logiques. On dit de ce modèle qu'il est « **boîte noire** ».

Il existe alors plusieurs indicateurs permettant de juger (abstraitemment et indictement) de la contribution d'une variable au modèle : la **Feature Importance** calculée selon la méthode **Gini Mean Decrease** [16].

Cette méthode se calcule de la manière suivante :

1. Calcul de l'erreur « Out of bag » du modèle complet
2. Calcul de l'erreur « Out of bag » du modèle dont la variable i a été retirée
3. soustraction de 2. par 1.

De cette manière, on aura un aperçu de la variable qui permet d'orienter au mieux le modèle vers la bonne prédiction puisque la différence sera maximale pour la variable qui collera le mieux à l'erreur OOB du modèle témoin.

Le graphe du Features Importance du modèle Random Forest qui vient corriger les erreurs de prédiction de la régression logistique mérite tout autant notre attention :

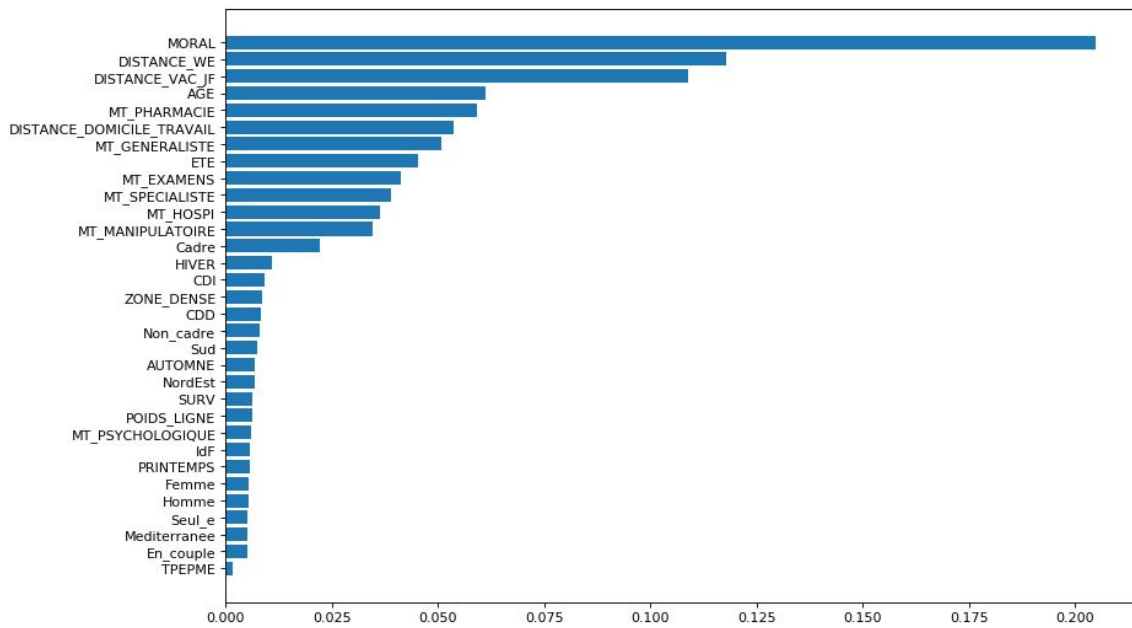


FIGURE 15 – Features Importance du Random Forest

En effet, le moral vient en première place en terme d'explicativité du modèle. Si l'on trace la courbe des survenances d'arrêt de travail par rapport à la courbe du moral que l'on a multiplié par -1 (car par hypothèse un moral faible devrait expliquer une forte survenance d'arrêts de travail et *vice versa* un moral élevé devrait justifier une faible survenance d'arrêts de travail).

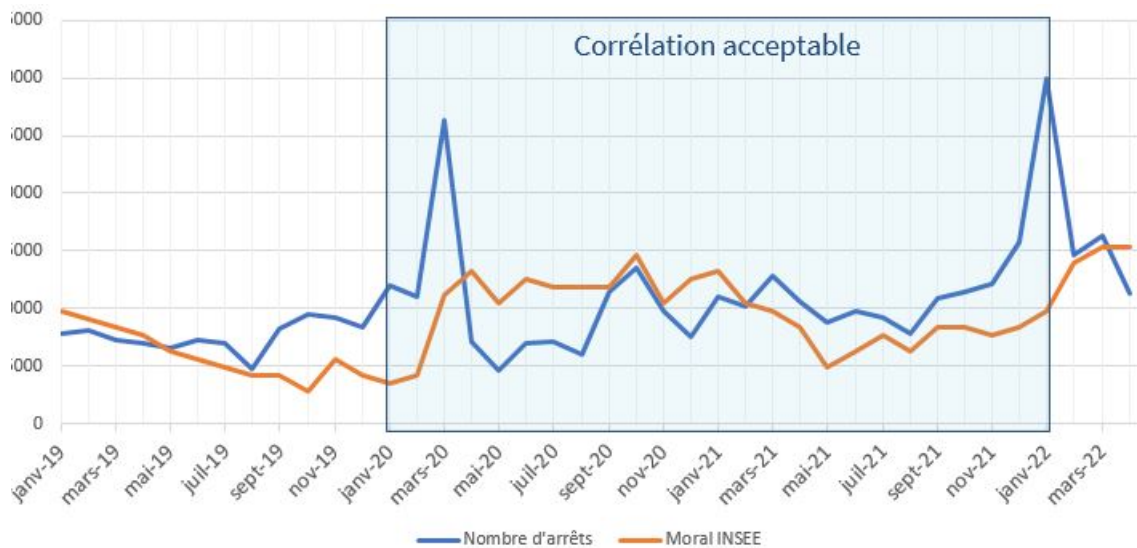


FIGURE 16 – Corrélation entre la survenance des arrêts de travail et l'opposé du moral INSEE

On a donc une tendance qui se détache et qui est la même pour les 2 courbes, notamment entre les mois de janvier 2020 et janvier 2022 (même si la corrélation est

positive sur le première semestre de 2019). Deux remarques :

1. On distingue des écarts non pas en direction mais en intensité sur les pics épidémiques de la première vague et de la cinquième vague (Omicron) de COVID-19, du fait des conditions d'isolement très strictes imposées à la population (confinement général resp. isolement des nombreux cas contacts). Le moral ne peut donc pas suivre la courbe des arrêts de travail avec un rapport de « 1 pour 1 ».
2. on observe aussi la fin de la corrélation après février 2022, le moral des français étant miné non plus par la pandémie mais par l'invasion de l'Ukraine par la Russie et la hausse du coût de la vie mais qui n'a logiquement pas d'impact sur l'absentéisme français.

Enfin, on apprend que le modèle Random Forest se sert en bonne partie des variables temporelles de distance aux jours vachés (vacances scolaires + jours fériés), du jour de la semaine et de l'âge.

On a donc un modèle « **2-en-1** » qui se sert d'une bonne partie des variables explicatives, la régression logistique se sert davantage des variables propres à l'individu alors que le Random Forest puise ses forces dans les variables macro propre à la survenance de l'élément perturbateur j .

Nous avons vu que le contexte temporel jouait un rôle non-négligeable dans la survenance d'un arrêt de travail (moral, proximité temporelle à un jour non-travaillé), mais aussi le temps de trajet et l'historique en arrêt maladie ; sera présenté dans la prochaine section ce qui peut caractériser la durée d'arrêt...

Prédiction de la durée d'un arrêt de travail

L'indicateur que nous voulons prédire s'exprimera comme la durée de l'arrêt, étant donné que c'est la seule donnée que l'on a.

La table d'étude comporte 585 583 lignes (les lignes comportant un 1 dans *SURV*).

Pour prédire cette variable continue, il nous faudra trouver un modèle de régression performant.

Voici la répartition des durées d'arrêt :

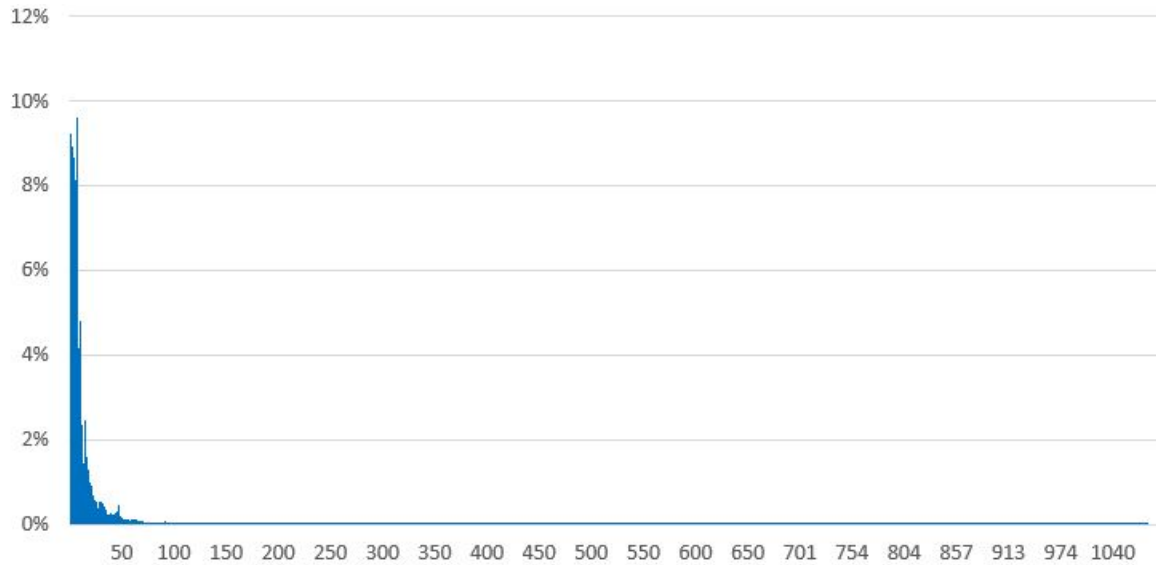


FIGURE 17 – Répartition de la durée d'arrêt

Ce qui donne en zoomant sur le gros de la distribution (la partie gauche) :

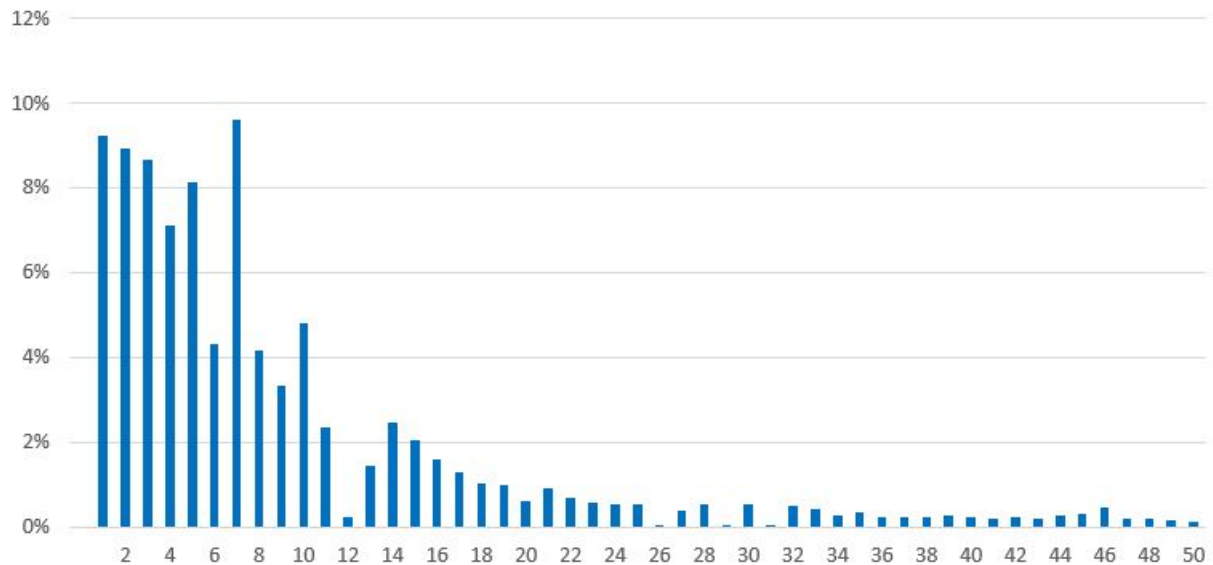


FIGURE 18 – Répartition de la durée d'arrêt sur $\llbracket 1, 50 \rrbracket$

On a donc un profil-type d'une loi de Poisson à quelques détails près... Les arrêts d'une semaine sont largement sur-représentés (durée = 5 jours si le début de l'arrêt est un lundi, 7 jours sinon), et à l'opposé les arrêts de 6 jours sont quasi-inexistants.

Pistes

L'une des difficultés de modélisation est liée à l'équilibre de notre table : 87% des arrêts durent moins de 30 jours (données > 1 an non affichées, 0.68% de la table) :

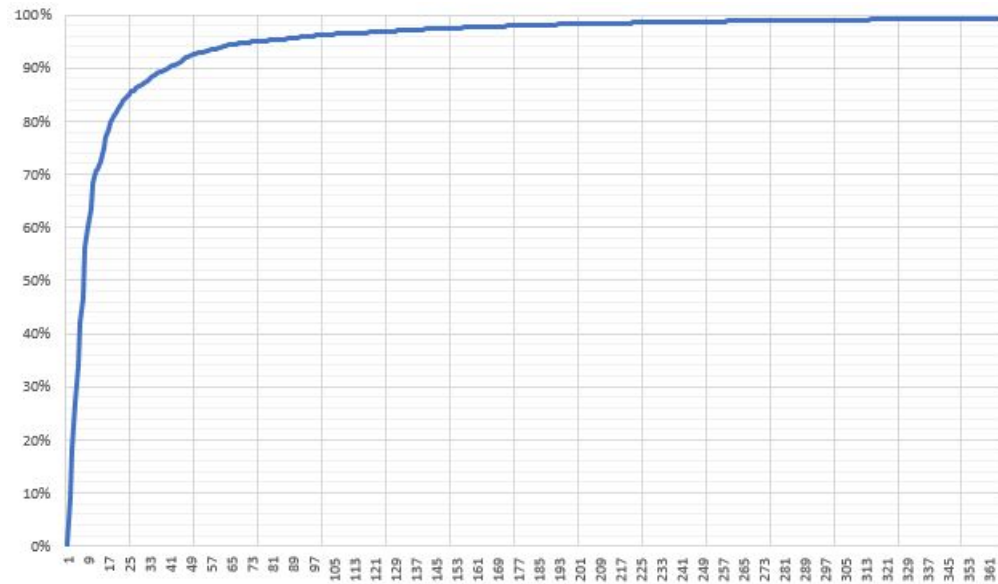


FIGURE 19 – Fonction de répartition de la durée d'arrêt

Nous allons donc séparer la durée d'arrêt en deux catégories et faire 2 modélisations différentes :

- Arrêts « courts », tous les arrêts d'une durée ≤ 30 jours
- Arrêts « longs », tous les arrêts d'une durée > 30 jours

Nous allons dans la section suivante modéliser la durée des **arrêts courts (moins d'1 mois)**...

Modélisation de la durée des arrêts de moins d'1 mois

GLM Poisson

Un premier GLM Poisson avec toutes les variables nous a indiqué devoir retirer les variables **secteur d'activité** et **consommation en pharmacie**. Voici le nouveau modèle sans ces deux variables :

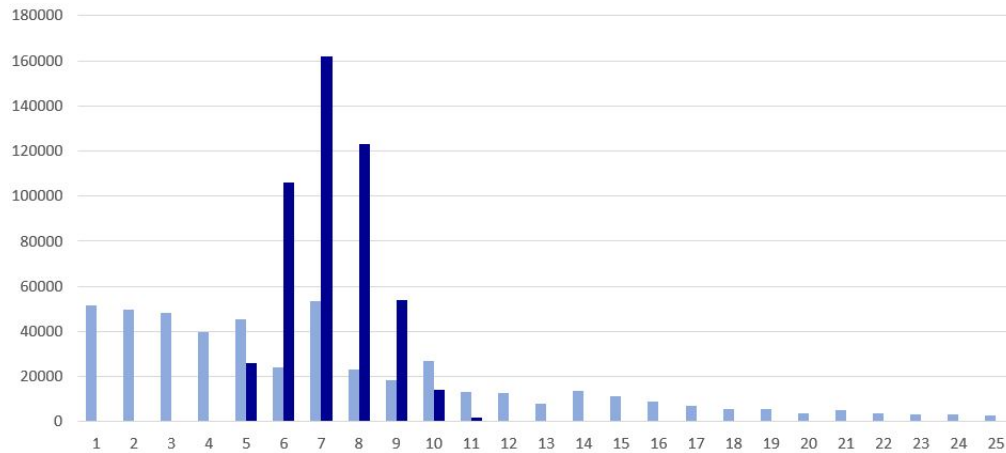


FIGURE 20 – Distribution du GLM Poisson (bleu foncé) vs la distribution réelle (bleu clair)

Voyons voir du côté des résidus à la maille individus :

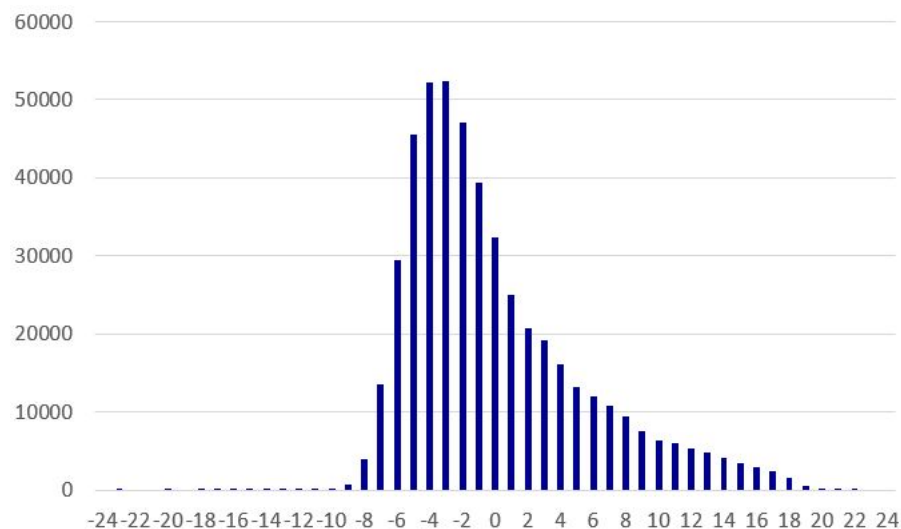


FIGURE 21 – Résidus de la modélisation par GLM Poisson à la maille salarié

Ces résidus sont à première vue mauvais, puisque la distribution n'est pas symétrique. Ce n'est pas très grave, car nous avons besoin des résidus à la maille entreprise (SIREN

et SIRET) pour notre étude : une prédiction tête par tête est bien trop compliquée et peu utile ; on devrait pouvoir aboutir au même résultat en agrégeant les résultats à une maille plus haute.

Voici la distribution des résidus moyens par **SIREN** et **SIRET** :

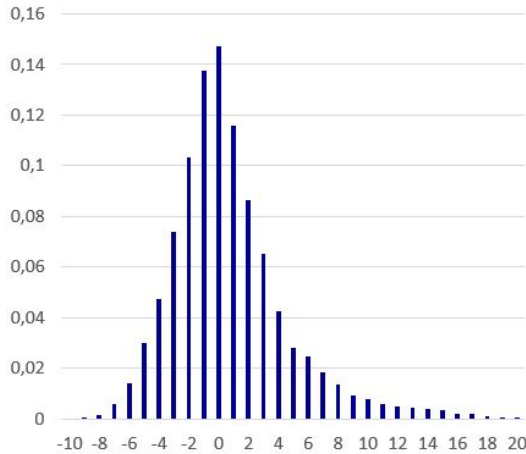


FIGURE 22 – Résidus de la modélisation par GLM Poisson à la maille SIREN

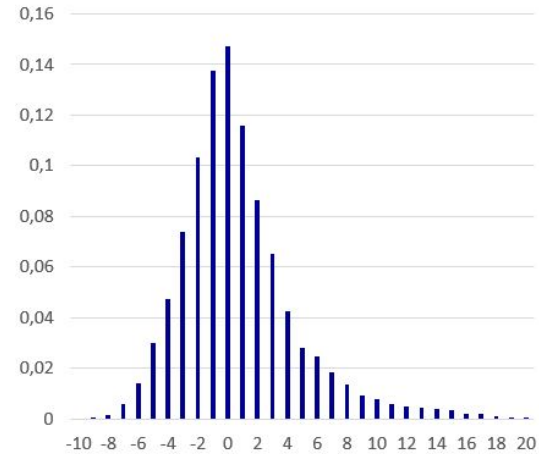


FIGURE 23 – Résidus de la modélisation par GLM Poisson à la maille SIRET

Nous pouvons vérifier la normalité de ces résidus par deux QQ-Plots :

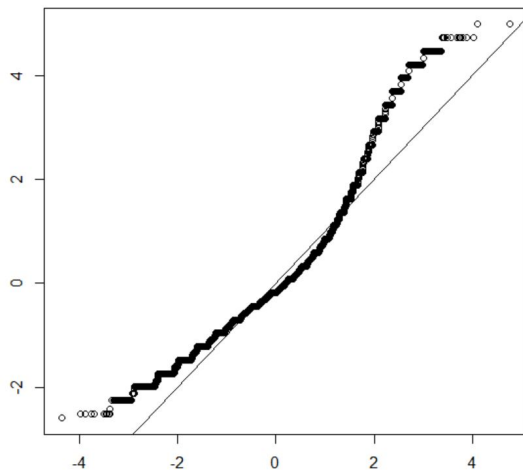


FIGURE 24 – QQ-Plot des résidus de la modélisation par GLM Poisson à la maille SIREN

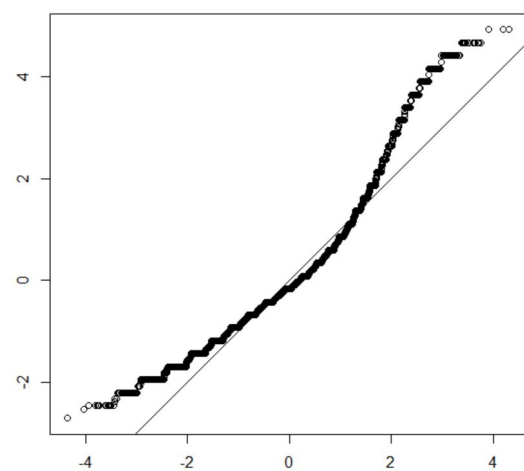


FIGURE 25 – QQ-Plot des résidus de la modélisation par GLM Poisson à la maille SIRET

On voit que les résidus ne sont pas normaux ; le test de normalité de **Shapiro-Wilk** viendra confirmer cette idée (**p-value** < **2e-16**, on rejette H_0 : « les données sont normalement distribuées ») ...

Comparons avec un **GLM Gamma**...

GLM Gamma

Le GLM Gamma donne à peu de choses près les mêmes résultats que le GLM Poisson :

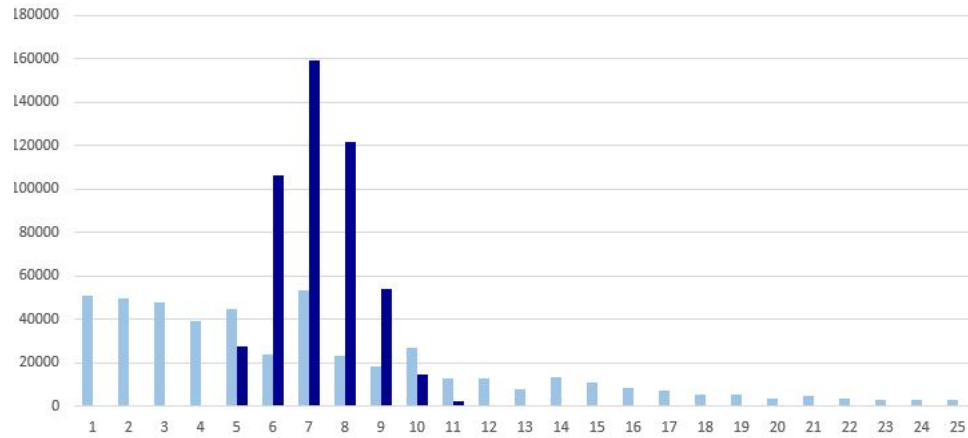


FIGURE 26 – Distribution du GLM Gamma (bleu foncé) vs la distribution réelle (bleu clair)

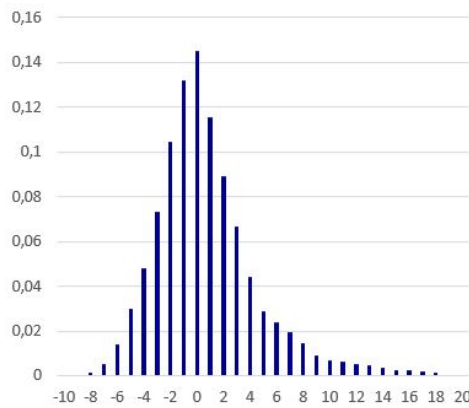


FIGURE 27 – Résidus de la modélisation par GLM Gamma à la maille SIREN

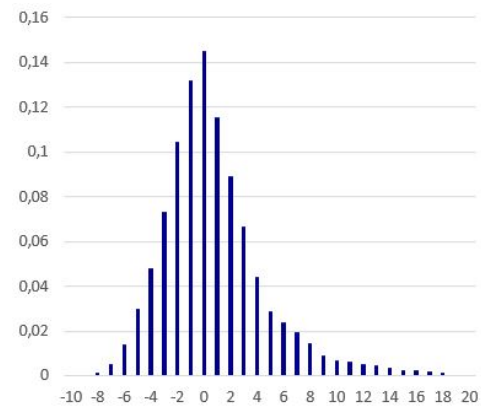


FIGURE 28 – Résidus de la modélisation par GLM Poisson à la maille SIRET

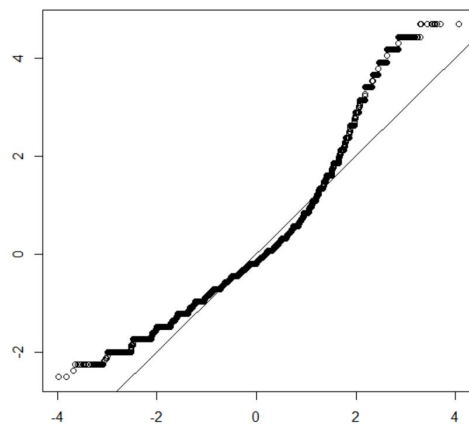


FIGURE 29 – QQ-Plot des résidus de la modélisation par GLM Gamma à la maille SIREN

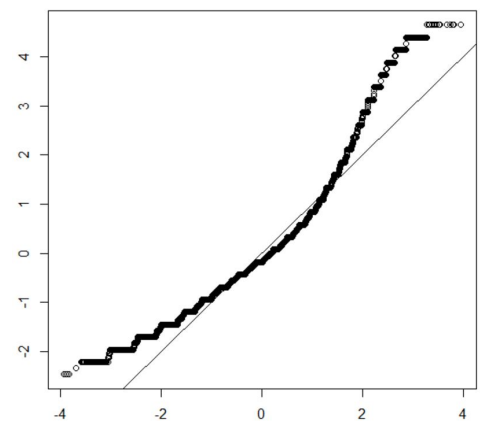


FIGURE 30 – QQ-Plot des résidus de la modélisation par GLM Gamma à la maille SIRET

Les GLM se montrent donc peu efficaces, essayons des modèles de Machine Learning...

Random Forest

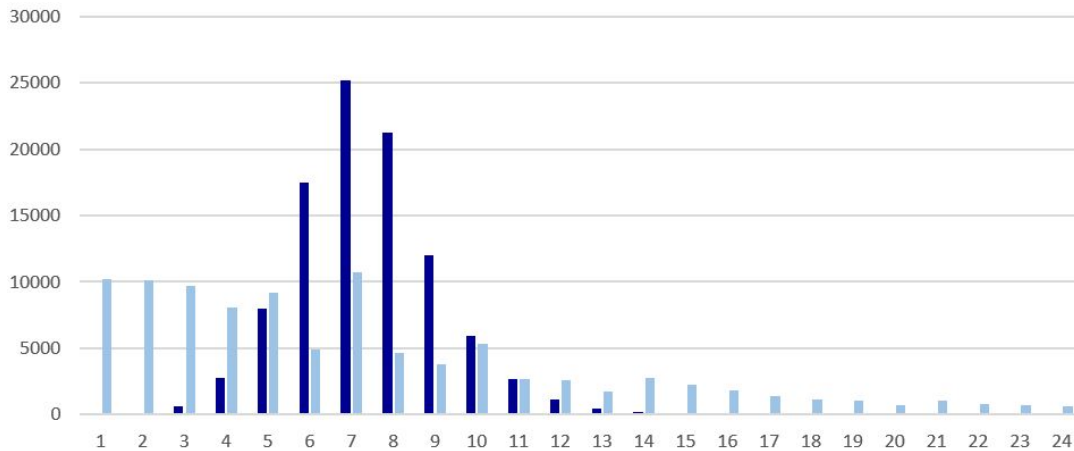


FIGURE 31 – Distribution du Random Forest (bleu foncé) vs la distribution réelle (bleu clair)

Le Random Forest nous propose une modélisation normale autour de 7 jours d'arrêt ce qui ressemble à peu près aux GLM, à ça près que l'algorithme est boîte noire... Nous allons essayer un XGBoost :

XGBoost (Extreme Gradient Boosting)

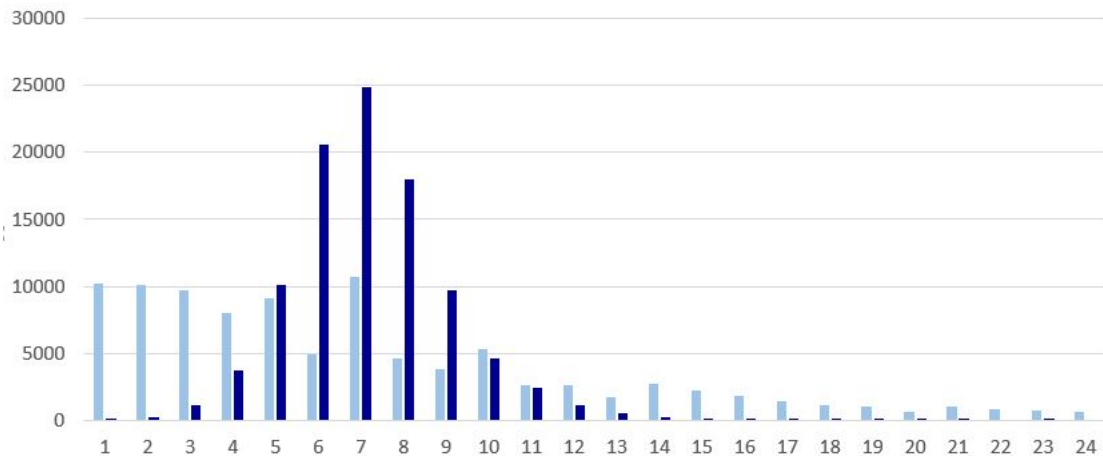


FIGURE 32 – Distribution du XGBoost (bleu foncé) vs la distribution réelle (bleu clair)

Le modèle XGBoost nous donne aussi cette même allure en terme de prédiction.

Récapitulatif

Nous nous retrouvons un peu dos au mur : aucun GLM ne fonctionne pour prédire efficacement la durée d'arrêt. De plus, les puissants modèles de Machine Learning que sont le **Random Forest** et le **XGBoost** ont beaucoup de mal à effectuer une régression pertinente.

Table de maintien en arrêt de travail

Notre idée est de créer une table de maintien en incapacité, où chaque colonne représenterait le taux de maintien d'un jour à l'autre (1 colonne = 1 jour). La somme de ces taux nous donnerait alors l'espérance de maintien en arrêt de travail, autrement dit un estimateur de la durée de l'arrêt de travail.

Dans le phénomène de l'interruption de travail, les individus sont hétérogènes entre eux. Cette hétérogénéité contient une part observable (ce que l'on a recensé dans la table : caractéristiques individuelles, consos en santé, historique d'inactivité) et une part déterministe non-observable. Dans ce contexte, l'étude du maintien en arrêt de travail ne peut être dissocié de ces caractéristiques hétérogènes.

La modélisation du risque de sortie d'arrêt peut être paramétrique, comme on vient de le voir : on choisit la loi de probabilité suivie par le risque de base. Un mauvais choix de cette loi aboutira à des estimateurs biaisés.

Une solution semi-paramétrique est de considérer l'intervalle de temps $[t-1, t[$ de sortie d'arrêt de travail. Ce modèle a l'avantage de ne pas imposer de loi, nous permettant de mieux coller à la réalité.

Pour ce faire, nous devons avoir la ventilation (combinaison de variables explicatives) la plus vaste possible tout en faisant en sorte que ces variables soient les plus discriminantes vis-à-vis de nos durées de maintien en arrêt de travail.

La **régression de Cox** [17] se présente comme le candidat favori pour connaître quelles variables retenir.

Le principe du modèle de Cox est de relier la date d'arrivée d'un événement à des variables explicatives. Par exemple, dans le domaine médical, on cherche à évaluer l'impact d'un prétraitement sur le temps de guérison d'un patient.

Il est considéré comme un modèle semi-paramétrique, et repose sur l'**hypothèse des risques proportionnels** :

Si l'on note h le risque instantané conditionnel aux covariables X_1, \dots, X_p :

$$h(t|X_{i,1}, \dots, X_{i,p}) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(T \in [t, t + dt] \mid T \geq t, X_{i,1}, \dots, X_{i,p})}{dt}$$

Alors l'hypothèse des risques proportionnels implique que :

$$h(t \mid X_{i,1}, \dots, X_{i,p}) = h_0(t) \exp(\theta_1 X_{i,1} + \dots + \theta_p X_{i,p}) = h_0(t) \exp(\theta X_i)$$

$h_0(t)$ est le risque de base, X_i est le vecteur des caractéristiques individuelles pour l'individu i . De ce fait, si X_1 est notre variable qualitative *Genre*,

$$X_{i,1} = \begin{cases} 1 & \text{si l'individu } i \text{ est un homme} \\ 0 & \text{si l'individu } i \text{ est une femme} \end{cases}$$

On doit avoir $\forall t$:

$$\frac{h_0(t) \exp(\theta_1 * 1 + \dots + \theta_p X_{i,p})}{h_0(t) \exp(\theta_1 * 0 + \dots + \theta_p X_{i,p})} = \exp(\theta_1)$$

$\exp(\theta_1)$ est le rapport de risque entre un homme et une femme toutes choses égales par ailleurs. On interprète la valeur de θ_1 ainsi :

θ_1	$\exp(\theta_1)$	Interprétation
< 0	< 1	maintien en arrêt plus long pour les hommes
> 0	> 1	maintien en arrêt plus long pour les femmes
$= 0$	$= 1$	maintien en arrêt équivalent pour femmes et hommes

Nous allons donc sélectionner nos variables d'intérêt selon si elles respectent cette hypothèse, d'autant plus que cela desssinera des tendances de risque instantané de tomber en arrêt de travail.

Voici un aperçu général (le *coef* est calculé par rapport à la modalité de référence pour chaque classe) :

n= 487532, number of events= 487532						
	coef	exp(coef)	se(coef)	z	Pr(> z)	
TR_AGE25 - 29	-2.622e-02	9.741e-01	6.283e-03	-4.173	3.01e-05	***
TR_AGE30 - 34	-8.330e-02	9.201e-01	6.327e-03	-13.165	< 2e-16	***
TR_AGE35 - 39	-1.505e-01	8.602e-01	6.521e-03	-23.086	< 2e-16	***
TR_AGE40 - 44	-1.933e-01	8.243e-01	6.802e-03	-28.411	< 2e-16	***
TR_AGE45 - 49	-2.530e-01	7.765e-01	6.986e-03	-36.211	< 2e-16	***
TR_AGE50 - 54	-2.793e-01	7.563e-01	7.149e-03	-39.074	< 2e-16	***
TR_AGE55 - 59	-3.121e-01	7.319e-01	7.440e-03	-41.945	< 2e-16	***
TR_AGE60 - 64	-3.642e-01	6.947e-01	1.017e-02	-35.825	< 2e-16	***
LIBELLE_CSPNonCadre	-1.508e-01	8.600e-01	3.973e-03	-37.951	< 2e-16	***
CONTRATCDI	-7.147e-02	9.310e-01	9.306e-03	-7.680	1.59e-14	***
ZONE_DENSE	8.548e-02	1.089e+00	3.441e-03	24.843	< 2e-16	***
DISTANCE_WE	7.406e-02	1.077e+00	1.533e-03	48.323	< 2e-16	***
SIT_FAMSeul_e	7.973e-03	1.008e+00	3.085e-03	2.585	0.00975	**

SAISONETE	-8.177e-02	9.215e-01	4.721e-03	-17.321	< 2e-16	***
SAISONHIVER	-4.839e-02	9.528e-01	3.784e-03	-12.787	< 2e-16	***
SAISONPRINTEMPS	-1.252e-01	8.824e-01	4.283e-03	-29.224	< 2e-16	***
DISTANCE_VAC_JF210 - 19	-3.897e-02	9.618e-01	3.168e-03	-12.302	< 2e-16	***
DISTANCE_VAC_JF220 - 29	1.025e-02	1.010e+00	6.488e-03	1.580	0.11418	
MT_PHARMACIE	9.719e-06	1.000e+00	3.486e-05	0.279	0.78039	
MT_HOSPI	-1.107e-04	9.999e-01	1.482e-05	-7.472	7.91e-14	***
MT_GENERALISTE	1.014e-04	1.000e+00	3.150e-05	3.218	0.00129	**
MT_SPECIALISTE	-5.345e-03	9.947e-01	4.032e-04	-13.256	< 2e-16	***
MT_EXAMENS	-1.350e-03	9.987e-01	1.516e-04	-8.906	< 2e-16	***
MT_MANIPULATOIRE	-1.046e-03	9.990e-01	2.461e-04	-4.249	2.15e-05	***
MT_PSYCHOLOGIQUE	-4.355e-05	1.000e+00	6.703e-05	-0.650	0.51583	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

D'après la dernière colonne, aucune de nos variables ne peut être retenue au titre de l'hypothèse des risques proportionnels sauf la consommation en pharmacie et psychologie, qui s'avèreront ne plus respecter l'hypothèse une fois seules dans une nouvelle modélisation par Cox.

Il existe certes une manière de respecter une hypothèse des risques proportionnels plus faible mais il y a bien trop de variables à prendre en compte.

Dans la prochaine partie sera proposée de créer une table de maintien en incapacité classique...

Calcul des taux de maintien

La probabilité de sortie de l'arrêt de travail lors du t^e jour s'écrit :

$$\mathbb{P}(\text{durée}(AT) \in [t, t+1]) = \mathbb{P}(\text{durée}(AT) \geq t) - \mathbb{P}(\text{durée}(AT) \geq t+1)$$

Et puisque la fonction de survie au début du t^e jour vaut :

$$S(t) = \mathbb{P}(\text{durée}(AT) \geq t)$$

On a donc que la probabilité de quitter l'arrêt lors du t^e jour vaut :

$$\mathbb{P}(\text{durée}(AT) \in [t, t+1]) = S(t) - S(t+1)$$

Et enfin, le taux de hasard du jour t (sortie d'arrêt de travail le t^e jour sachant que l'on n'était pas sorti avant) vaut :

$$h(t \mid X_{i,1}, \dots, X_{i,p}) = \frac{S(t) - S(t+1)}{S(t)} = 1 - \frac{S(t+1)}{S(t)} \quad (1)$$

Estimateur de Kaplan-Meier

Les méthodes statistiques couramment utilisées pour construire une table de maintien reposent sur l'estimateur de Kaplan-Meier [18] [19] :

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{d_i}{n_i}\right)$$

où \hat{S} représente l'estimateur de la fonction de survie, d_i représente le nombre de personnes terminant un arrêt de travail au jour t_i et n_i le nombre de personnes encore en arrêt à l'instant t_{i-} .

Si l'on revient à l'équation (1), puisque l'on peut estimer S à l'aide de Kaplan-Meier par \hat{S} , (1) devient :

$$\hat{h}(t \mid X_{i,1}, \dots, X_{i,p}) = 1 - \frac{\hat{S}(t+1)}{\hat{S}(t)} = \frac{d_t}{n_t}$$

Construction de la table de maintien ventilée

La table de maintien reprendra le format de celle du **BCAC**. C'est une table à deux dimensions :

- Une correspond à la maille d'entrée en arrêt de travail (âge, CSP, contrat, situation familiale, agglomération, proximité au week-end et aux jours vauqués, saison)
- l'autre représente l'ancienneté dans l'arrêt de travail (ici en **jours**, entre 1 et 30)

Cette table contient dans un premier temps les $l_{x,i}$, à partir desquelles on peut déduire la majorité des grandeurs actuarielles (notamment nos taux).

Ceci étant fait, nous pouvons créer notre table de maintien qui comporte d'abord les $l_{x,i}$, puis calculer les taux de maintien d'un jour sur l'autre.

Cette table a donc cette forme :

i	GENRE	TR AGE	CSP	Contrat	GdeVille	WE	Sit familiale	JV	0	1	...	30
1	F	20-24	Cadre	CDI	Oui	2j	En couple	0-9j	39	36	...	0
2	F	20-24	Cadre	CDI	Oui	2j	En couple	10-19j	25	22	...	0
3	F	20-24	Cadre	CDI	Oui	2j	En couple	20-29j	14	14	...	0
4	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	0-9j	35	32	...	0
5	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	10-19j	24	22	...	0
6	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	20-29j	16	14	...	0
...
fin	H	60-64	N.Cadre	CDD	Non	3j	Seul(e)	20-29j	26	22	...	0

Nous pouvons alors calculer la probabilité de se maintenir du jour t au jour $t + 1$:

$$p_t = 1 - q_t = 1 - \frac{d_t}{l_t}$$

Cette formule peut être répétée à volonté, pour n jours de maintien :

$${}_n d_t = d_t + \dots + d_{t+n-1} = l_t - l_{t+n}$$

avec ${}_n d_t$ le nombre de personnes terminant un arrêt de travail entre le jour t et le jour $t + n$. On a donc :

$${}_n q_t = \frac{{}_n d_t}{l_t} \text{ et } {}_n p_t = \frac{l_{t+n}}{l_t}$$

Nous allons donc remplir la partie droite du tableau avec nos ${}_n p_t$, plus précisément nos ${}_n p_0$ avec $n \in \{1, \dots, 30\}$:

i	GENRE	TR AGE	CSP	Contrat	GdeVille	WE	Sit familiale	JV	0	1	...	30
1	F	20-24	Cadre	CDI	Oui	2j	En couple	0-9j	1	0,92	...	0
2	F	20-24	Cadre	CDI	Oui	2j	En couple	10-19j	1	0,88	...	0
3	F	20-24	Cadre	CDI	Oui	2j	En couple	20-29j	1	1	...	0
4	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	0-9j	1	0,91	...	0
5	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	10-19j	1	0,92	...	0
6	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	20-29j	1	0,88	...	0
...
fin	H	60-64	N.Cadre	CDD	Non	3j	Seul(e)	20-29j	1	0,85	...	0

L'espérance de maintien peut donc se calculer de la forme suivante :

$$e_x = \sum_{k=1}^{+\infty} {}_k p_x$$

Dans notre cas, nous voulons l'espérance de maintien de la ligne i à partir de la survenance (jour $x = 0$) :

$$E_{\text{maintien},i} = e_{0,i} = \sum_{k=1}^{30} {}_k p_{0,i} = \frac{1}{l_{0,i}} \sum_{k=1}^{30} l_{k,i}$$

Ce qui donne (valeurs fictives) :

i	GENRE	TR AGE	CSP	Contrat	GdeVille	WE	Sit familiale	JV	0	1	...	30	E
1	F	20-24	Cadre	CDI	Oui	2j	En couple	0-9j	39	0,92	...	0	7
2	F	20-24	Cadre	CDI	Oui	2j	En couple	10-19j	25	0,88	...	0	6
3	F	20-24	Cadre	CDI	Oui	2j	En couple	20-29j	14	1	...	0	10
4	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	0-9j	35	0,91	...	0	5
5	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	10-19j	24	0,92	...	0	6
6	F	20-24	Cadre	CDI	Oui	2j	Seul(e)	20-29j	16	0,88	...	0	7
...
fin	H	60-64	N.Cadre	CDD	Non	3j	Seul(e)	20-29j	26	0,85	...	0	3

Cette table de maintien ventilée va donc servir à projeter des durées d'arrêt en la joignant à notre table d'étude...

Rapatriement des prédictions par la table de maintien en incapacité sur notre table d'étude

Le travail de cette partie va consister à joindre la table des prédictions de durées en fonction du **genre, de la tranche d'âge, de la CSP, du contrat de travail, de la présence en agglomération, de la proximité au week-end et aux jours fériés, et de la situation familiale**, colonnes que l'on retrouve également sur notre table d'étude. Cette étape va nous permettre de confronter les données prédites aux données réelles, afin d'évaluer la **précision** et la **pertinence** de ce modèle.

En rapatriant les prédictions des arrêts courts sur notre table d'étude, les durées d'arrêt sont réparties selon une distribution qui a l'allure suivante :

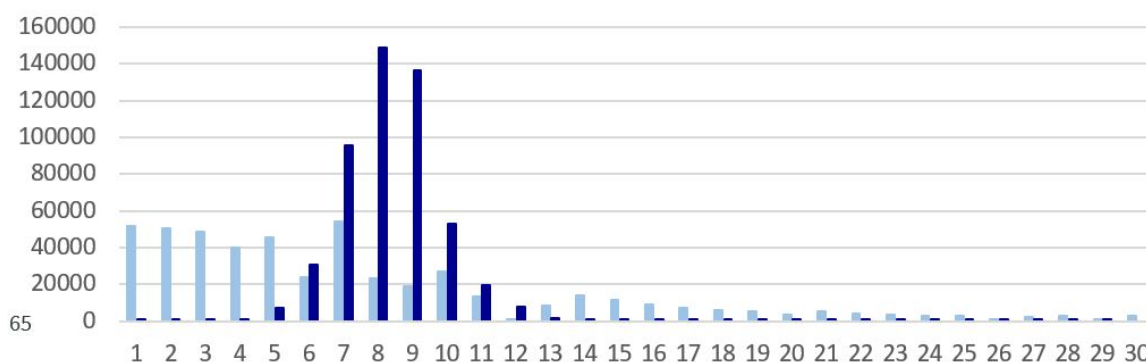


FIGURE 33 – Répartition des durées d'arrêt de travail réelles (bleu clair) et prédites par la table de maintien en incapacité (bleu foncé)

Nous sommes encore une fois confrontés à une distribution normale autour de 7 jours, Il semblerait ainsi que les méthodes de **Machine Learning** soient aussi efficaces qu'une méthode actuarielle. Pour nous aider à choisir, nous allons regarder ce que donnent les moyennes des résidus moyens par SIREN et SIRET pour (certains de) nos modèles :

	GLM Poisson	GLM Gamma	Table de maintien
Résidu moyen par SIREN	0,72j	0,74j	0,47j
Résidu moyen par SIRET	0,75j	0,68j	0,38j

Conclusion : On peut donc choisir d'utiliser la table de maintien pour modéliser nos durées d'arrêts de moins d'1 mois.

Modélisation de la durée des arrêts > 30 jours

Pour modéliser ces durées élevées et très disparates, nous avons réfléchi à plusieurs pistes pour le type de la variable cible. Faut-il garder les durées d'arrêt sous un format continu ? Ou devons-nous séparer cette variable en classes ?

Durée d'arrêt de travail quantitative

Si la variable réponse **durée de l'arrêt** restait quantitative, voici les modèles possibles :

- Quantitatif #1 : GLM : Poisson, Gamma
- Quantitatif #2 : Valeurs extrêmes avec une GPD
- Quantitatif #3 : Machine Learning
- Quantitatif #4 : Table de maintien en incapacité prenant pour origine une durée de maintien égale à 1 mois

Les 3 premières options ont donné des résultats trop éloignés de la réalité. En effet, les modélisations ont été *stricto sensu* continues. La forte dispersion et la répartition assez aléatoire des arrêts très longs n'a pas été respectée.

De plus, les options 2 et 3 sont boîte noire ce qui ne permet pas une analyse pertinente.

Durée d'arrêt de travail qualitative

Une autre piste a été de convertir la durée d'arrêt en classes : 1-2mois, 2-3 mois, 3-6 mois, 6 mois-1an, 1 an et +. Ensuite seront appliqués nos algorithmes de classification par *Machine Learning*. Le problème est le même : les classes sont trop déséquilibrées et les classes les plus « hautes » ne sont jamais bien prédites.

On se rapatrie donc sur la **table de maintien en incapacité** décalée d'1 mois (option #4), cette fois-ci en appliquant des taux de SMR sur certaines modalités discriminantes...

Table de maintien en incapacité supérieure à 1 mois

Notre table de maintien en incapacité « longue » sera construite selon des critères différents par rapport à ceux utilisés pour la précédente table de maintien (arrêts « courts »).

On ne va plus chercher à ventiler nos taux de maintien par chaque variable explicative mais cette étape sera dorénavant réalisée après un calcul unique par tranche d'âge : on

appliquera ensuite un coefficient propre à chaque modalité de chaque variable explicative ; ces coefficients sont appelés **Standardized Mortality Ratio** (SMR).

Explications :

1. on calcule une table de maintien très similaire à celle du BCAC : en lignes les âges entre 20 et 65 ans, et en colonnes les mois de 1 à 36. On calcule les taux de la manière classique (l_{x+t}/l_x) puis on somme pour obtenir le maintien en mois (à multiplier par 30, et translater de 30 car le maintien d'1 mois est une hypothèse).
2. Ensuite nous calculons les rapports SMR qui pour chaque variable explicative souhaitée s'exprime comme :

$$SMR_i = \frac{\text{Durée d'arrêt totale prédite pour la cohorte } i}{\text{Durée d'arrêt totale réelle pour la cohorte } i}$$

3. Nous obtenons les taux SMR suivants :

Variable	Modalité	Taux SMR
Genre	Homme	1,022
	Femme	0,973
CSP	Cadre	1,066
	Non cadre	0,99
Contrat	CDI	0,996
	CDD	1,332
Situation familiale	En couple	1,025
	Seul(e)	0,976

Un taux supérieur à 1 signifie pour la cohorte que le maintien de la population générale (numérateur) est plus long que le maintien réel de cette cohorte (dénominateur). De ce fait, la durée de maintien estimée doit être divisée par ce taux de SMR pour coller à la durée de maintien théorique. Nous avons donc des durées de maintien ajustées à la maille **genre x CSP x type de contrat x situation familiale** pour nos arrêts dits « longs ». Les tendances qui se détachent de ces calculs sont donc les suivantes :

- les hommes se maintiennent en arrêt moins longtemps que les femmes toutes choses égales par ailleurs
- les cadres se maintiennent en arrêt moins longtemps que les non-cadres toutes choses égales par ailleurs
- les CDD se maintiennent en arrêt moins longtemps que les CDI toutes choses égales par ailleurs
- les personnes en couple se maintiennent en arrêt moins longtemps que les personnes seules toutes choses égales par ailleurs

Concaténation des 2 tables de maintien en arrêt « court » et « long » avec leurs prédictions

Afin d'établir la précision du modèle final pour la durée d'arrêt (« court » + « long »), nous allons afficher les mêmes données qu'avant, sur notre table entière : résidus moyen par SIREN et par SIRET :

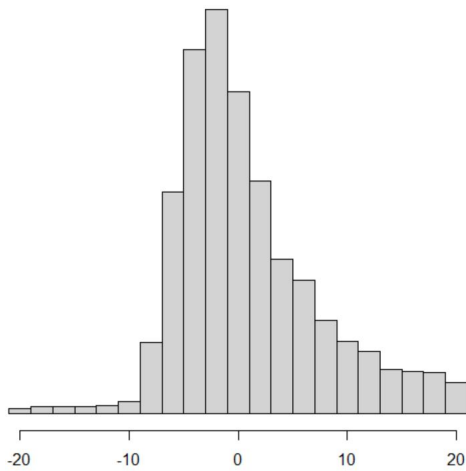


FIGURE 34 – Résidus des prédictions de durée d'arrêt à la maille SIREN

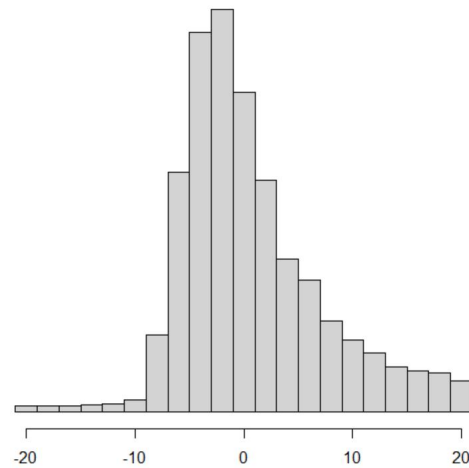


FIGURE 35 – Résidus des prédictions de durée d'arrêt à la maille SIRET

De plus, nos erreurs sont certes tirées vers le haut (en valeur absolue) par les arrêts très longs (au-delà d'1 an), mais restent décentes :

	Maintien arrêts courts	Maintien arrêts longs	Global
Résidu moyen par SIREN	0,47j	-8,84j	-1,61j
Résidu moyen par SIRET	0,75j	-9,39j	-1,8j

Voilà à peu près tout pour les modèles, pour comprendre quelle variable tire la durée d'arrêt dans tel sens.

Dans le prochain chapitre sera abordée la stratégie à porter par le groupe, que ce soit en interne pour suivre ce risque, le piloter, le cerner (notamment tirer les premières conclusions de l'impact Covid) ou pour conseiller les clients Collectives afin de leur proposer d'anticiper cet absentéisme en fonction de leurs effectifs, mais aussi de réduire cet absentéisme avec des solutions adaptées à chaque situation...

Comment maîtriser cet absentéisme au regard de cette étude ?

Généralités sur l'absentéisme

Certaines définitions de l'absentéisme insistent sur le caractère délibéré, d'autres sur la récurrence de l'arrêt de travail, ou encore sur son caractère imprévisible. L'**absentéisme** caractérise toute absence qui aurait pu être évitée par une analyse et une prévention précoce des facteurs de dégradation des conditions de travail.

Il a plusieurs causes, notamment le vieillissement de la population active, et l'élévation du taux d'activité.

Ces absences ont plusieurs impacts nuisibles à l'entreprise et aux autres salariés, d'après l'**Agence Nationale pour l'Amélioration des Conditions de Travail (ANACT)** :

- productivité moindre, qualité de service dégradée, expérience client amputée
- réorganisation des effectifs de travail
- répartition souvent disproportionnée de la charge de travail sur les salariés présents, nourrissant le sentiment de faire le travail des absents
- dégradation des indicateurs sociaux, baisse de motivation

L'analyse de ses causes peut permettre de traiter le mal à la racine, d'autant plus que c'est dans l'intérêt conjoint du salarié et de l'entreprise.

Les coûts de l'absentéisme

Évaluer l'absentéisme en entreprise est assez facile si l'on se cantonne à la seule indemnisation des salariés, mais mesurer son impact sur toute la chaîne de valeur demeure nettement plus compliqué pour les raisons évoquées au paragraphe précédent. On estime au total que l'absentéisme coûterait plus de **100 milliards d'euros** chaque an aux entreprises. [3]

Les coûts directs de l'absentéisme

Ces coûts sont les plus quantifiables, puisqu'ils représentent la part de salaire de l'employé à maintenir pendant l'arrêt de travail. De plus, ces coûts restent modérés puisque la Sécurité Sociale se charge d'une bonne partie du maintien de salaire. Ceux-ci dépendent de la politique RH de l'entreprise et de la convention collective ou accord national interprofessionnel.

Les coûts indirects de l'absentéisme

Certains coûts doivent aussi être assumés par l'entreprise en cas d'absence de l'un de ses salariés. Il s'agit de coûts internes pour assurer une continuité de service [20]. Ces coûts sont principalement :

- Les **coûts de gestion de l'absence** : charge supplémentaire pour les RH
- Les **coûts de remplacement** : salarié remplaçant à payer

Les coûts cachés de l'absentéisme

Enfin, il y a des coûts qui sont très difficiles à quantifier tant il y a de facteurs qui peuvent les impacter [20].

- Les **coûts d'image pour l'entreprise** : service non entièrement assuré, insatisfaction client
- Les **coûts sociaux** : lassitude des autres salariés, notamment ceux sur qui la charge de travail retombe. Risque d'effet boule de neige.
- Les **coûts de perte de productivité** : le remplacement d'un salarié ne sera pas aussi efficace que si le salarié « originel » travaillait.

En effet, ces coûts peuvent grandement altérer la qualité de service, ce qui peut retomber sur l'entreprise en terme d'image. Enfin, l'absentéisme risque de dégrader les conditions de travail des salariés présents ce qui peut déboucher sur d'autres arrêts de travail.

Ces coûts cachés peuvent prendre une ampleur conséquente si les cadences de travail sont élevées, si l'effectif est restreint, si les climats sociaux local et national (cf moral INSEE) sont tendus...

L'évolution de la situation ces dernières années

Le taux d'absentéisme est la grandeur qui permet de mesurer à une maille plus large (SIRET, SIREN, secteur, région, pays, etc.) la situation. La formule est la suivante :

$$T_{\text{absentéisme}} = \frac{\text{nombre total de jours de travail manqués}}{\text{nombre total de jours théoriquement travaillés}}$$

Ces 2 dernières années, le taux d'absentéisme a énormément crû du fait de la crise sanitaire, comme l'atteste l'histogramme suivant [4] :

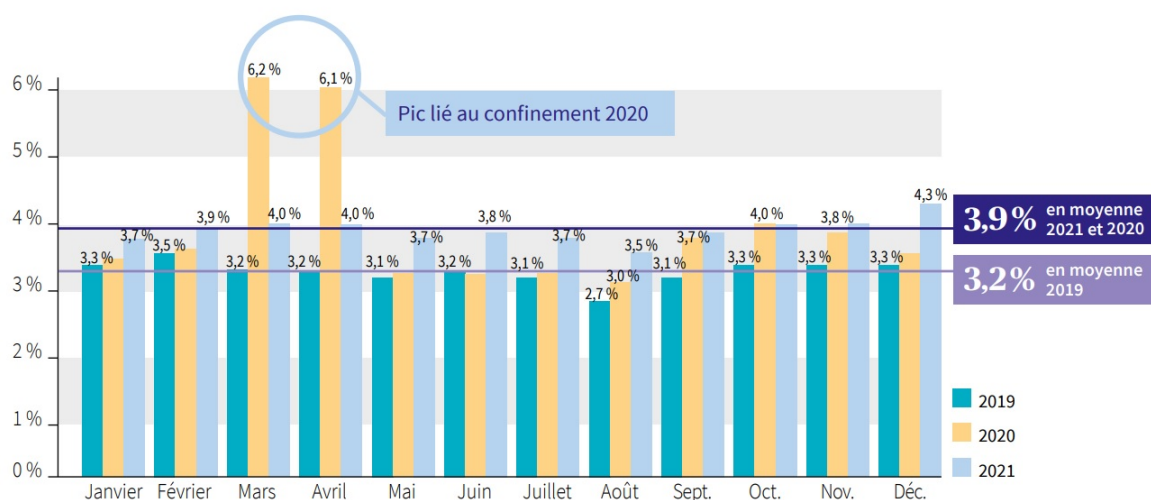


FIGURE 36 – Evolution du taux d'absentéisme pour les entreprises couvertes par AXA

Nous assistons donc à une dégradation de cet indicateur avec la crise sanitaire. Si l'on zoome par durée d'absence :

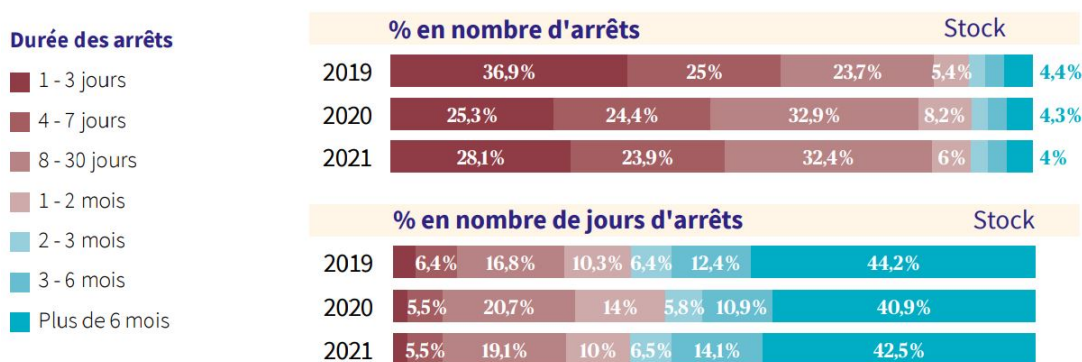


FIGURE 37 – Répartition du nombre et de la durée des absences pour les entreprises couvertes par AXA

Nous voyons sur ces graphes que les arrêts d'une durée supérieure à 3 mois sont d'un nombre plutôt stable mais leur durée s'allonge notablement ; les **risques psychosociaux** en représentent la première cause.

Nous allons donc voir dans le prochain chapitre que non seulement la prévention est la solution la plus instinctive mais aussi la plus efficace.

La prévention comme l'une des solutions

Les bilans de santé

Bilans de santé physique

Une bonne partie des arrêts longs sont causés par des **Affections Longue Durée** (ALD). Ces arrêts longs ont non seulement des effets dévastateurs pour les salariés, mais ce sont aussi les arrêts les plus coûteux pour l'entreprise et la Sécurité Sociale. On retrouve notamment (liste non-exhaustive) :

- les **cancers** (382 000 cas par an)
- les **accidents cardio-vasculaires invalidants** (150 000 AVC par an)
- le **diabète** (4 millions de français)
- la **maladie d'Alzheimer** (225 000 cas par an)
- les **hépatites** (150 000 français)
- le **VIH** et le **SIDA** (6 000 cas par an)
- la **maladie de Parkinson** (25 000 cas par an)
- la **sclérose en plaques** (3 000 cas par an)

Bilans de santé mentale

Concernant l'aspect de la santé mentale, tous les voyants sont au rouge depuis la crise sanitaire. On estime par exemple que **25% des français étaient anxieux** en juin 2022 [21]. Or on a vu qu'il y avait une forte corrélation entre le **moral des ménages** synthétisé par l'INSEE et la survenance d'un arrêt de travail grâce au Random Forest.

De ce fait, il est légitime de penser qu'ouvrir des canaux d'écoute psychologique permettra de déceler des premiers signes d'affaiblissement de la santé mentale, de mesurer l'ampleur éventuelle de la souffrance pour fournir un accompagnement psychologique adapté. Celui-ci peut passer par des consultations avec un psychologue mis à la disposition exclusive de l'entreprise, qui pourra par exemple prescrire un arrêt de travail en prévention et dont la durée sera modérée plutôt que de laisser le mal-être progresser.

Un **diagnostic précoce** de ces maladies permet de commencer un traitement qui stopperait la progression de la maladie. C'est donc un pari gagnant-gagnant. Il faut adopter une stratégie de bilans de santé physique et mentale réguliers et exhaustifs, pour pouvoir démarrer un traitement rapidement si nécessaire...

Miser sur la prévention et les traitements précoces pour éviter la propagation de cet absentéisme

Toutes les **ALD** citées ci-dessus ainsi que les troubles psychologiques doivent être diagnostiqués et traités le plus tôt possible. En effet, certaines de ces pathologies peuvent évoluer en dépendance partielle ou totale. Les salariés requèrent parfois un accompagnement quotidien en fonction de l'état de dépendance dans lequel ils se trouvent. Dans ce cas, trois possibilités s'offrent à ces personnes :

- le placement en établissement spécialisé
- le maintien à domicile avec passage d'un aide-soignant
- le maintien à domicile avec aide d'un proche

Les deux premières solutions étant très **onéreuses**, la troisième solution se retrouve souvent choisie. Si bien que l'on estime qu'entre **8 et 11 millions de Français** sont aidants [22] : conjoints, enfants, proches, etc. Ces personnes peuvent donc requérir de s'absenter (pour un motif légitime) afin d'aider ledit proche.

Ces proches-aidants comme on les appelle font l'objet d'un absentéisme plus élevé que la moyenne et 9 sur 10 sont sujets au stress [23]. Cette tâche qui leur incombe les pousse par exemple à :

- arriver plus tard ou partir plus tôt
- poser des jours de congé à la dernière minute
- être moins impliqué au travail

Ainsi, la prévention est essentielle pour diminuer cet absentéisme contagieux.

La prévention est un levier très puissant, mais il faut évidemment rendre le travail plus attractif pour éviter l'absentéisme non-causé par des pathologies, ce que l'on verra dans la section suivante...

Améliorer les conditions de travail pour améliorer l'engagement

Diminuer le mal-être au travail

Le mal-être au travail est le principal coupable dans cet absentéisme élevé. En effet, certains salariés peuvent décider de se mettre en arrêt de manière abusive pour esquiver le travail. Parmi les raisons qui peuvent expliquer ce geste, on retrouve :

- un manque de reconnaissance
- des horaires trop larges
- un conflit interne
- trop de distance entre le domicile et le lieu de travail (par exemple pour garder ou récupérer un enfant)

Cette incitation à ne pas solliciter le médecin pour des problèmes mineurs peut permettre de réduire les arrêts courts, dits de complaisance. Pour y remédier, on pense notamment à :

- la **contre-visite médicale**
- la **visite au domicile**

Si l'arrêt de travail s'avère avoir été prescrit de manière abusive, alors les indemnités touchées seront retirées.

Cependant, la répression ne fera pas grandement diminuer l'absentéisme. Il faudrait donc privilégier une amélioration de la qualité de vie au travail (QVT) qui passe par, par exemple [4] [5] :

- un meilleur équilibre vie personnelle - vie professionnelle
- plus d'autonomie (télétravail)
- des perspectives d'évolution, une rémunération attractive
- un accompagnement pour le retour après un arrêt
- plus de communication *bottom-up*
- soutenir les « toujours présents »

Ces solutions sont donc exclusivement à mettre en place du côté de l'employeur car il s'agit de politiques RH.

Nous allons enfin voir dans la partie suivante une stratégie concrète de réduction de l'absentéisme et leurs impacts sur diverses entreprises.

Proposition de plan de réduction de l'absentéisme et simulation de son impact sur le taux d'absentéisme

Cette section qui clôturera le mémoire vient synthétiser toutes les mesures évoquées ci-dessus pour diminuer le taux d'absentéisme. Un tel scénario sera appliqué sur 3 entreprises ayant des populations différentes, et sur le portefeuille complet.

Présentation des entreprises témoin

En plus de **notre table de travail**, nous allons sélectionner 3 entreprises aux profils divers :

- **Entreprise A**, la grande entreprise au plus fort taux d'absentéisme
- **Entreprise B**, l'entreprise à la plus forte démographie
- **Entreprise C**, l'entreprise au plus fort chiffre d'affaires

Entreprise A



FIGURE 38 – Présentation de l'entreprise A

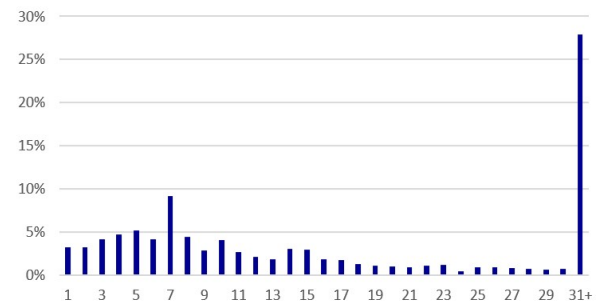


FIGURE 39 – Répartition des arrêts de l'entreprise A selon la durée en jours

Entreprise B



FIGURE 40 – Présentation de l'entreprise B

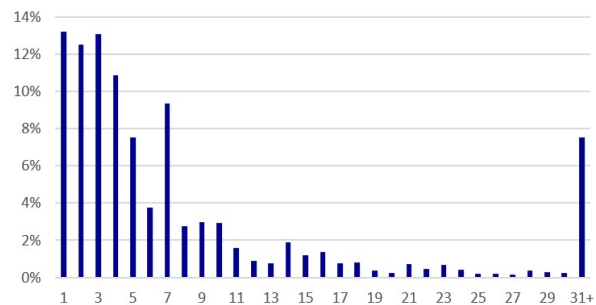


FIGURE 41 – Répartition des arrêts de l'entreprise B selon la durée en jours

Entreprise C



FIGURE 42 – Présentation de l'entreprise C

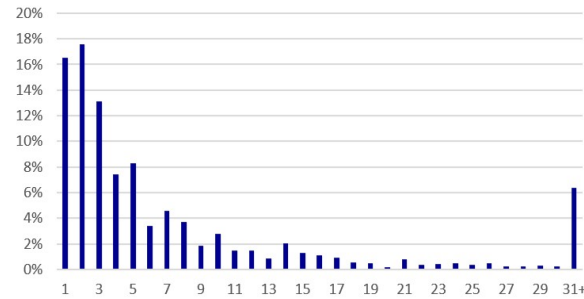


FIGURE 43 – Répartition des arrêts de l'entreprise C selon la durée en jours

Périmètre global



FIGURE 44 – Présentation du périmètre global

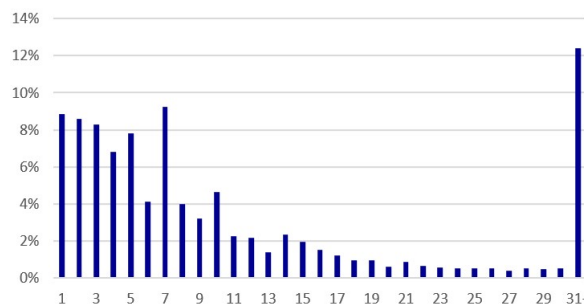


FIGURE 45 – Répartition des arrêts du périmètre global selon la durée en jours

Ci-dessous nous allons voir quelles solutions AXA peut apporter et dans quelle mesure cela ferait baisser le taux d'absentéisme...

Hypothèses

Hypothèse de prise en charge précoce des ALD

On estime que le VIH est dépisté à un stade avancé dans 28% des cas. Il n'y a pas de chiffre pour tous les cancers, uniquement pour le cancer du poulmon : 50% des dépistages le sont trop tardivement.

Comme discuté plus tôt, nous allons fixer comme objectif que 20% des complications dues à une ALD sont évitables grâce à une prise en charge précoce. L'idée est donc d'aller chercher les arrêts maladie longs ayant des combinaisons de consommations en

santé sur les postes **Examens** ou **Hospitalisation** pour réduire leur nombre de manière stochastique de 20%.

Hypothèse de prise en charge précoce des troubles psychologiques

Concernant le volet psychologique, 33% des personnes interrogées lors de la vague 34 de CoviPrev [21] étaient sujets à un état dépressif et/ou anxieux. Ces valeurs n'ont que très peu bougé depuis le premier confinement, et sont plus élevées de 14 points par rapport à la situation pré-Covid.

Si l'on se fixe pour scénario une réduction de moitié de cet écart entre pré et post-Covid, on suppose que **7/33 = 21%** des arrêts de travail ayant un profil de consommation en psychologue sont évitables.

Hypothèse d'amélioration des conditions de travail

Dans une optique d'améliorer les conditions de travail, plus de flexibilité doit être accordée au salarié. On pense en premier lieu au télétravail, avec la libre décision des jours qui permettrait par exemple de s'occuper de son enfant malade, mais aussi qui éviterait le trajet domicile-travail souvent éreintant.

Nous n'avons pas de donnée concernant la part de salariés absents pour cause de « trajet trop long » mais nous pouvons supposer que 15% des arrêts de travail qui ont lieu en zone dense et avec une distance domicile-travail de plus de 25 kilomètres sont évitables avec du télétravail (ce qui fait minimum 45 minutes de trajet), ou grâce à un aménagement des horaires. Nous allons appliquer cette décôte aux seuls Cadres, puisqu'ils disposent d'un métier souvent télétravaillable.

Enfin, 36% des salariés estiment que leur travail a un impact significatif sur leur état de santé. Si nous souhaitons descendre ce taux à 30%, nous allons éliminer de manière aléatoire 6% des arrêts de travail en fin de chaîne.

Résultats de la modélisation

Nous allons alors appliquer l'algorithme suivant :

1. **Au titre des ALD** : retirer 20% des arrêts au hasard si l'une des conditions est remplie :
 - le montant de consommation en examens est supérieur au quantile d'ordre 0,8
 - le montant de consommation en hospitalisation est supérieur au quantile d'ordre 0,8 et l'arrêt dépasse 1 mois
2. **Au titre de la santé psychologique** : retirer 21% des arrêts au hasard avec une consultation chez psychologue
3. **Au titre du télétravail** : retirer 15% des arrêts au hasard si le profil est un cadre habitant en zone dense et travaillant à + de 25 kilomètres de son domicile
4. **Au titre de l'amélioration des conditions de travail en général** : retirer 6% des arrêts au hasard sur la table restante

Chacune des étapes ci-dessus générera un périmètre sur lequel le taux d'absentéisme sera calculé. Cet algorithme sera répété **250 fois** puis la moyenne sera prise pour obtenir des valeurs stables et non soumises à des tirages qui risqueraient de biaiser l'étude.

Voici ces taux d'absentéisme, mesurés par des impacts **cumulée** :

Taux absentéisme (%)	Entreprise A	Entreprise B	Entreprise C	Global
Base	7,58	2,20	2,18	3,69
Étape 1	6,89	1,97	1,92	3,31
Étapes 1+2	6,79	1,90	1,86	3,26
Étapes 1+2+3	6,77	1,85	1,84	3,26
Étapes 1+2+3+4	6,37	1,74	1,73	3,06

Ce qui donne en impact pour une étape isolée :

Évol. taux d'abs (%)	Entreprise A	Entreprise B	Entreprise C	Global
Impact Étape 1	-10	-11,7	-13,5	-11,5
Impact Étape 2	-1,5	-3,7	-3,2	-1,5
Impact Étape 3	-0,3	-2,7	-1,1	0
Impact Étape 4	-6,3	-6,3	-6,4	-6,5
Impact Total	-19	-26,4	-26	-20,6

À première vue, chaque étape contribue différemment à faire baisser le taux d'absentéisme selon l'entreprise à laquelle elle est appliquée. On remarque par exemple que :

- l'impact de la « **prévention des ALD** » réduit d'environ 11% le taux d'absentéisme mais ce chiffre est plus élevé pour l'entreprise C, probablement car la prévention et les traitements sont coûteux et donc accessibles aux classes sociales les plus aisées. Ces disparités sociales biaisent sans doute notre étude puisque les classes sociales les moins aisées peuvent ne pas consulter pour une ALD mais en être affecté tout de même.
- l'impact de la « **prévention des troubles psychologiques** » réduit d'environ 1,5% le taux d'absentéisme mais ce chiffre est relativement plus élevé pour les entreprises B et C qui sont des CCN avec des métiers de bureau et dont les employés sont de ce fait davantage sujets au *burn-out*.
- l'impact du « **télétravail** » aurait un impact plus puissant sur une entreprise comptant davantage de cadres qui peuvent souvent gérer leur travail à distance de par leurs tâches

Nous voyons donc que ces pistes de réduction de l'absentéisme peuvent réduire de 20 à 25% le taux d'absentéisme. Cela est évidemment sous réserve d'atteindre les objectifs fixés dans les hypothèses, mais aussi que les trois parties (salarié, entreprise, assureur) soient des parties prenantes à cet effort.

Conclusion

L'absentéisme est un sujet intéressant car au-delà de la dimension actuarielle, il apporte une bonne compréhension de la société française. L'analyse des données permet d'éclairer les connaissances déjà acquises sur ce sujet et de voir que de nouvelles tendances dans le comportement des salariés ressortent pendant et au lendemain de la crise sanitaire.

L'étude a requis avant toute chose de comprendre que l'absentéisme pouvait être multifactoriel et a nécessité un long travail de réflexion, de cohérence pour créer notre table d'étude. Ce travail est une étape *sine qua non* de l'étude.

Dans un premier temps, nous avons compris à l'aide d'un double modèle régression logistique et Random Forest que la survenance d'un arrêt était corrélée à la consultation d'un médecin généraliste, à la proximité temporelle avec un jour non-travaillé, à l'historique d'arrêts maladie, à la taille de l'entreprise, à la présence en zone congestionnée et enfin (beaucoup) au moral des ménages. Cela confirme le ressenti majeur depuis la crise sanitaire consistant en une forte demande de meilleure équilibre vie professionnelle/vie personnelle.

Ensuite, l'estimation de la durée de l'arrêt de travail à l'aide des consommations de santé et des variables macro s'est montrée inopérante puisque tout profil de consommation de santé correspond à toute durée d'arrêt de travail (exemple : un arrêt peut durer 1 an sans qu'il y ait de consommation explicite auparavant, tout comme un arrêt peut durer 1 an pour *burn-out* avec une trace de consommation). Nous avons donc certains pans de la population ayant un maintien en arrêt plus long à savoir : les femmes (+2,7%), les non-cadres (+1%), les personnes en CDI (+0,4%), les personnes seules (+2,4%). Le fort déséquilibre de la table au profit des arrêts de durée courte ou modérée nous a contraint à être moins précis sur la modélisation des arrêts longs, l'objectif était alors de perdre le moins d'information sur les arrêts courts. Cependant, une étude concentrée uniquement sur les arrêts longs peut être pertinente (valeurs très extrêmes). Un algorithme SMOTE peut être à l'ordre du jour pour ré-échantillonner ces longs arrêts.

En dernier temps ont été développées les causes des arrêts de travail au regard de l'étude et de la réalité du terrain. Une modélisation a été proposée en agissant sur 4 leviers que sont :

1. la prévention et le traitement des Affections Longue Durée
2. la détection des signes de mal-être psychologique
3. la généralisation du télétravail
4. l'amélioration plus globale des conditions de travail pour recréer de l'engagement

Ces 4 efforts tripartites si réalisés à 100% des objectifs fixés, pourraient faire baisser l'absentéisme de plus de 20% ce qui ferait jusqu'à 25 milliards d'euros de plus pour l'économie et donc probablement reversés en tant que hausses de salaire.

Remerciements

Dans un premier temps, je tiens à remercier **Fabienne Cazals**, responsable de l'équipe Etudes et Innovations Data, qui m'a permis de faire mon alternance au sein de la Direction Data d'**AXA Santé & Collectives**. Sa bonne humeur et sa gestion a favorisé mon intégration dans cette équipe jeune et riche de talents. Par cette occasion je souhaite également remercier l'ensemble des collaborateurs de l'équipe pour leur accueil, leur aide et plus globalement pour leur présence.

Comme discuté dans ce mémoire, les bonnes conditions de travail permettent d'améliorer les performances, et une collaboration efficace est le fruit d'une excellente cohésion d'équipe.

Ensuite, je veux remercier plus particulièrement **Vincent Decamps**, mon tuteur en entreprise dans un premier temps, et **Déborah Hulot**, qui a pris le relais pour m'encadrer : ils m'ont accompagné, épaulé, formé et fait progresser en me proposant des missions et des axes d'étude pertinents. Leur regard sur ce sujet qui a pris une importance inédite avec l'arrivée des DSN et l'irruption de la crise sanitaire a été précieux. Ils ont fait preuve de bienveillance et de disponibilité à mon égard.

J'ai pu apprendre auprès d'eux les métiers d'Actuaire et Data Scientist, les attentes, les compétences requises, les outils, les méthodes de travail, l'arbitrage, l'autonomie, la polyvalence, la richesse des missions et des domaines d'expertise.

Maintenant, je voudrais remercier **Olivier Lopez**, mon tuteur académique et aussi directeur de l'ISUP. Plus généralement, je veux remercier les **équipes pédagogiques de l'ISUP** pour la qualité de leur enseignement, notamment pendant la crise sanitaire.

Je veux fermer ce chapitre en adressant ma plus profonde gratitude à **mes parents** qui m'ont toujours soutenu au cours de mes études.

Bibliographie

- [1] ONU, *La santé mentale au travail*, 28 Septembre 2022, URL : <https://www.who.int/fr/news-room/fact-sheets/detail/mental-health-at-work>
- [2] URSSAF, Samarin, *La déclaration sociale nominative (DSN)*, URL : <https://www.urssaf.fr/portail/DSN>
- [3] Institut Sapiens, *Le coût caché de l'absentéisme au travail*, Novembre 2018, URL : <https://www.institutsapiens.fr/wp-content/uploads/2018/11/Cout-absenteisme.pdf>
- [4] AXA, *Absentéisme et prévention*, 2022, URL : <https://www.axa-assurancescollectives.fr/wp-content/uploads/2022/06/Barometre-absenteisme-prevention-edition-2022.pdf>
- [5] AG2R La Mondiale, *14ème Baromètre de l'Absentéisme et de l'Engagement*, 2022, URL : <https://www.ag2rlamondiale.fr/files/live/sites/portail/files/pdf/Culture-branches/Barom%C3%A8tres%20Ayming/14eme-barometre-ayming-ALM-2022.pdf>
- [6] Legifrance, *Loi EVIN*, 1989, URL : <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000709057>
- [7] Cindy CORNUAILLE, *La prévoyance collective*, Septembre 2020
- [8] URSSAF, *La portabilité des garanties de prévoyance*, URL : <https://www.urssaf.fr/portail/home/employeur/calculer-les-cotisations/les-elements-a-prendre-en-compte/la-prevoyance-complementaire/la-portabilite-des-garanties-de.html>
- [9] Legifrance, *Code de la Sécurité Sociale*, Novembre 2022, URL : <https://www.legifrance.gouv.fr/codes/id/LEGITEXT000006073189/>
- [10] Wikipedia, *Formule de Haversine*, URL : https://fr.wikipedia.org/wiki/Formule_de_haversine
- [11] TomTom, *France Traffic*, 2022, URL : <https://www.tomtom.com/traffic-index/france-country-traffic>
- [12] INSEE, *Moral des ménages*, Avril 2022, URL : <https://www.insee.fr/fr/statistiques/6436843>

- [13] Charlotte Dion, *Cours de modèle linéaire à l'ISUP*, 2021
- [14] Claire Boyer, *Cours de Machine Learning à l'ISUP*, 2021
- [15] Marina Kia pour jedha.co, *Qu'est-ce que la cross-validation ?*, URL : <https://www.jedha.co/formation-ia/cross-validation>
- [16] StackExchange, *How to interpret Mean Decrease in Accuracy and Mean Decrease GINI in Random Forest models*, Novembre 2018, URL : <https://stats.stackexchange.com/questions/197827/how-to-interpret-mean-decrease-in-accuracy-and-mean-decrease-gini-in-random-fore>
- [17] Olivier Bouaziz (Paris Descartes), *Analyse de survie : le modèle de Cox*, URL : <https://helios2.mi.parisdescartes.fr/obouaziz/CoxSurv.pdf>
- [18] Olivier Lopez, *Cours de modèle de durée à l'ISUP*, 2022
- [19] Frédéric Planchet, *Modèles de durée, arrêt de travail*, 2020-2021, URL : [http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/1430AD6748CE3AFFC1256F130067B88E/\\$FILE/Seance7.pdf?OpenElement](http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/1430AD6748CE3AFFC1256F130067B88E/$FILE/Seance7.pdf?OpenElement)
- [20] Damien Péan pour gereso.com, *Absentéisme : comment calculer son coût réel ?*, Mai 2018, URL : <https://www.gereso.com/actualites/2018/06/05/absenteisme-comment-calculer-son-cout-reel/>
- [21] Santé Publique France, *Comment évolue la santé mentale des Français pendant l'épidémie de COVID-19 – Résultats de la vague 34 de l'enquête CoviPrev*, Septembre 2022, URL : <https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-et-infections-respiratoires/infection-a-coronavirus/documents/enquetes-etudes/comment-evolue-la-sante-mentale-des-francais-pendant-l-epidemie-de-covid-19-resultats-de-la-vague-34-de-l-enquete-coviprev>
- [22] OCIRP, *Les chiffres-clés sur les aidants en France*, Octobre 2018, URL : <https://ocirp.fr/actualites/les-chiffres-cles-sur-les-aidants-en-france>
- [23] Essentiel Autonomie / Malakoff Humanis, *Employeurs : quelles difficultés pour vos salariés aidants ?*, Novembre 2021, URL : <https://www.essentiel-autonomie.com/concilier-aidance-travail/employeurs-queelles-difficultes-vos-salaries-aidants>

Annexes

Annexe 1 : Sortie de la régression logistique avec toutes les variables pour prédire la survenance d'un arrêt de travail

(Intercept)	1.807e+01	3.559e+05	0.000	1.0000
MT_PHARMACIE	1.692e-05	1.644e-04	0.103	0.9180
MT_HOSPI	3.042e-03	1.664e-04	18.282	< 2e-16 ***
MT_GENERALISTE	3.722e-02	1.088e-03	34.221	< 2e-16 ***
MT_SPECIALISTE	1.491e-03	7.982e-04	1.868	0.0617 .
MT_EXAMENS	2.926e-04	1.558e-04	1.878	0.0603 .
MT_MANIPULATOIRE	-9.732e-03	6.923e-04	-14.058	< 2e-16 ***
MT_PSYCHOLOGIQUE	-4.998e-03	9.330e-04	-5.357	8.45e-08 ***
AGE	-9.816e-03	4.037e-04	-24.317	< 2e-16 ***
SEXEM	-2.038e-01	9.754e-03	-20.900	< 2e-16 ***
SIT_FAMSeul_e	-2.198e-01	9.560e-03	-22.993	< 2e-16 ***
LIBELLE_CSPNonCadre	2.594e-01	1.186e-02	21.878	< 2e-16 ***
SECTEURAGRO-ALIMENTAIRE	-2.603e+01	3.559e+05	0.000	0.9999
SECTEURAUTRES	-2.642e+01	3.559e+05	0.000	0.9999
SECTEURBANQUES, ETABLISSEMENTS FINANCIERS ET ASSURANCES	-2.587e+01	3.559e+05	0.000	0.9999
SECTEURBATIMENTS ET TRAVAUX PUBLICS	-2.590e+01	3.559e+05	0.000	0.9999
SECTEURBOIS ET DERIVES	-2.603e+01	3.559e+05	0.000	0.9999
SECTEURBRANCHES AGRICOLES	-2.621e+01	3.559e+05	0.000	0.9999
SECTEURBUREAUX D'ETUDES ET PRESTATIONS DE SERVICES AUX ENTREPRISES	-2.638e+01	3.559e+05	0.000	0.9999
SECTEURCHIMIE ET PHARMACIE	-2.590e+01	3.559e+05	0.000	0.9999
SECTEURCOMMERCE DE DETAIL PRINCIPALEMENT NON ALIMENTAIRE	-2.631e+01	3.559e+05	0.000	0.9999
SECTEURCOMMERCE DE GROS ET IMPORT-EXPORT	-2.599e+01	3.559e+05	0.000	0.9999
SECTEURCOMMERCE PRINCIPALEMENT ALIMENTAIRE	-2.591e+01	3.559e+05	0.000	0.9999
SECTEURCULTURE, SPORT, MEDIA ET COMMUNICATION	-2.638e+01	3.559e+05	0.000	0.9999
SECTEUREAU ET ENERGIE	-2.600e+01	3.559e+05	0.000	0.9999
SECTEURENSEIGNEMENT ET FORMATION	-2.638e+01	3.559e+05	0.000	0.9999
SECTEURHABILLEMENT, CUIR, TEXTILE	-2.643e+01	3.559e+05	0.000	0.9999
SECTEURHOTELLERIE, RESTAURATION ET TOURISME	-2.641e+01	3.559e+05	0.000	0.9999
SECTEURIMMOBILIER ET ACTIVITES TERTIAIRES LIEES AU BÂTIMENT	-2.621e+01	3.559e+05	0.000	0.9999
SECTEURMETALLURGIE ET SIDERURGIE	-2.578e+01	3.559e+05	0.000	0.9999
SECTEURNETTOYAGE, MANUTENTION, RECUPERATION ET SECURITE	-2.642e+01	3.559e+05	0.000	0.9999
SECTEURPLASTIQUES, CAOUTCHOUC ET COMBUSTIBLES	-2.587e+01	3.559e+05	0.000	0.9999
SECTEURPROFESSIONS JURIDIQUES ET COMPTABLES	-2.635e+01	3.559e+05	0.000	0.9999
SECTEURSECTEUR SANITAIRE ET SOCIAL	-2.614e+01	3.559e+05	0.000	0.9999
SECTEURSERVICES DE L'AUTOMOBILE ET DES MATERIELS ROULANTS	-2.605e+01	3.559e+05	0.000	0.9999
SECTEURTRANSPORT ET LOGISTIQUE	-2.621e+01	3.559e+05	0.000	0.9999
SECTEURVERRES ET MATERIAUX DE CONSTRUCTION	-2.586e+01	3.559e+05	0.000	0.9999
SECTEURVRP	-2.659e+01	3.559e+05	0.000	0.9999
TAILLE_ENTGE	5.558e-01	3.490e-02	15.927	< 2e-16 ***
TAILLE_ENTTPE/PME	-1.567e-01	1.172e-02	-13.369	< 2e-16 ***
DISTANCE_DOMICILE_TRAVAIL	-8.835e-04	4.315e-05	-20.477	< 2e-16 ***
ZONE_DENSE	-2.131e-01	1.303e-02	-16.355	< 2e-16 ***
MORAL	6.587e-02	9.523e-04	69.164	< 2e-16 ***

DISTANCE_VAC_JF	6.720e-02	6.943e-04	96.783	< 2e-16	***
DISTANCE_WE	4.702e-01	4.847e-03	96.998	< 2e-16	***
SAISONETE	-1.475e-01	1.627e-02	-9.062	< 2e-16	***
SAISONHIVER	1.067e-01	1.343e-02	7.945	1.94e-15	***
SAISONPRINTEMPS	-7.093e-01	1.473e-02	-48.137	< 2e-16	***
SEV_MALADIE	1.223e+04	2.719e+03	4.497	6.89e-06	***
SEV_ATMP	3.934e+03	5.226e+03	0.753	0.4515	
SEV_TPP	7.109e+02	1.771e+03	0.401	0.6882	
CONTRATCDI	1.421e+00	2.101e-02	67.635	< 2e-16	***
RegionMediterranee	1.557e-01	1.490e-02	10.450	< 2e-16	***
RegionNordEst	1.772e-01	1.652e-02	10.729	< 2e-16	***
RegionNordOuest	1.723e-01	1.608e-02	10.715	< 2e-16	***
RegionSud	1.984e-01	1.467e-02	13.525	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 518223 on 635082 degrees of freedom
Residual deviance: 307930 on 635027 degrees of freedom
AIC: 299213

Number of Fisher Scoring iterations: 25

Annexe 2 : Sortie de la régression logistique avec les variables significatives pour prédire la survenance d'un arrêt de travail

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-3.9337	-0.3554	0.0000	0.3627	3.6168

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-8.083e+00	9.770e-02	-82.737	< 2e-16 ***
MT_HOSPI	3.078e-03	1.659e-04	18.548	< 2e-16 ***
MT_GENERALISTE	3.724e-02	1.082e-03	34.429	< 2e-16 ***
MT_MANIPULATOIRE	-9.363e-03	6.741e-04	-13.888	< 2e-16 ***
MT_PSYCHOLOGIQUE	-4.738e-03	9.172e-04	-5.166	2.39e-07 ***
AGE	-8.454e-03	3.900e-04	-21.676	< 2e-16 ***
SEXEM	-1.325e-01	8.979e-03	-14.760	< 2e-16 ***
SIT_FAMSeul_e	-2.445e-01	9.415e-03	-25.972	< 2e-16 ***
LIBELLE_CSPNonCadre	3.291e-01	1.086e-02	30.289	< 2e-16 ***
TAILLE_ENTGE	4.963e-01	3.414e-02	14.535	< 2e-16 ***
TAILLE_ENTTPE/PME	-1.613e-01	1.132e-02	-14.254	< 2e-16 ***
DISTANCE_DOMICILE_TRAVAIL	-9.723e-04	4.241e-05	-22.928	< 2e-16 ***
ZONE_DENSE	-3.382e-01	1.240e-02	-27.263	< 2e-16 ***
MORAL	6.534e-02	9.391e-04	69.581	< 2e-16 ***
DISTANCE_VAC_JF	6.697e-02	6.876e-04	97.388	< 2e-16 ***

DISTANCE_WE	4.693e-01	4.781e-03	98.159	< 2e-16	***
SAISONETE	-1.525e-01	1.605e-02	-9.499	< 2e-16	***
SAISONHIVER	9.259e-02	1.328e-02	6.975	3.06e-12	***
SAISONPRINTEMPS	-7.190e-01	1.452e-02	-49.521	< 2e-16	***
SEV_MALADIE	1.238e+04	2.723e+03	4.546	5.46e-06	***
CONTRATCDI	1.405e+00	2.051e-02	68.500	< 2e-16	***
RegionMediterranee	1.466e-01	1.466e-02	10.006	< 2e-16	***
RegionNordEst	1.798e-01	1.624e-02	11.071	< 2e-16	***
RegionNordOuest	1.717e-01	1.579e-02	10.873	< 2e-16	***
RegionSud	2.037e-01	1.443e-02	14.116	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 518223 on 635082 degrees of freedom
 Residual deviance: 315080 on 635058 degrees of freedom
 AIC: 305347

Number of Fisher Scoring iterations: 25

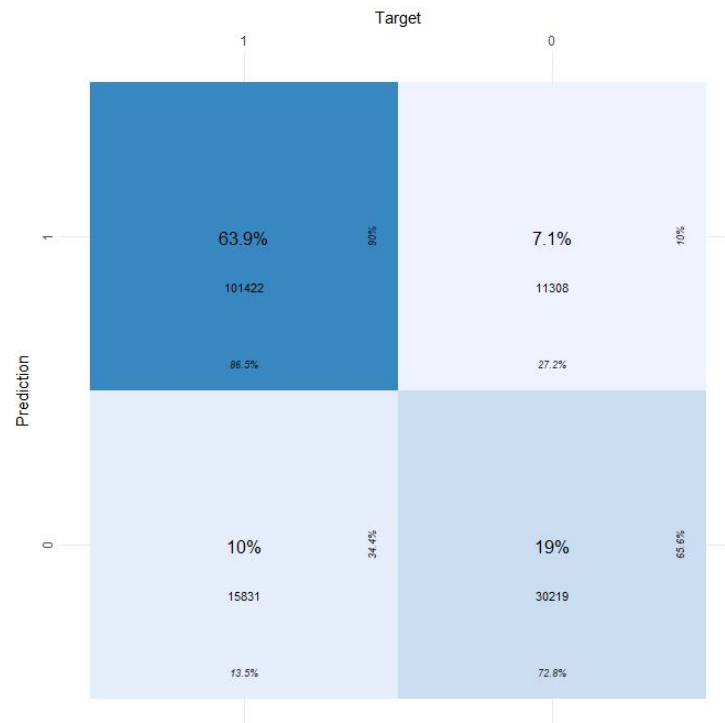


FIGURE 46 – Matrice de confusion de la régression logistique + bayésien naïf calibré à 0.5

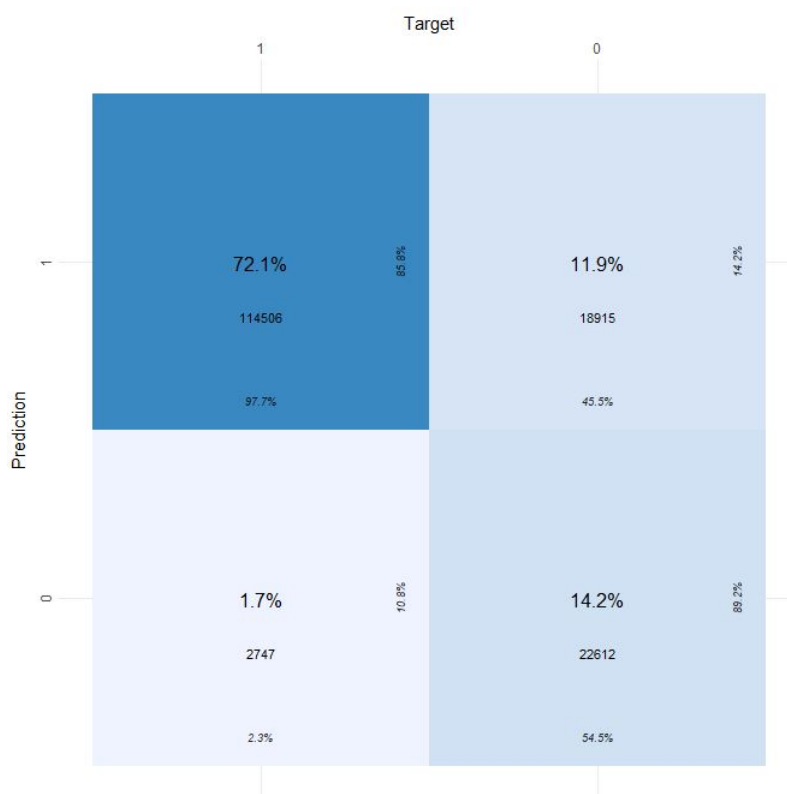


FIGURE 47 – Matrice de confusion de la régression logistique + bayésien naïf calibré à 0.26

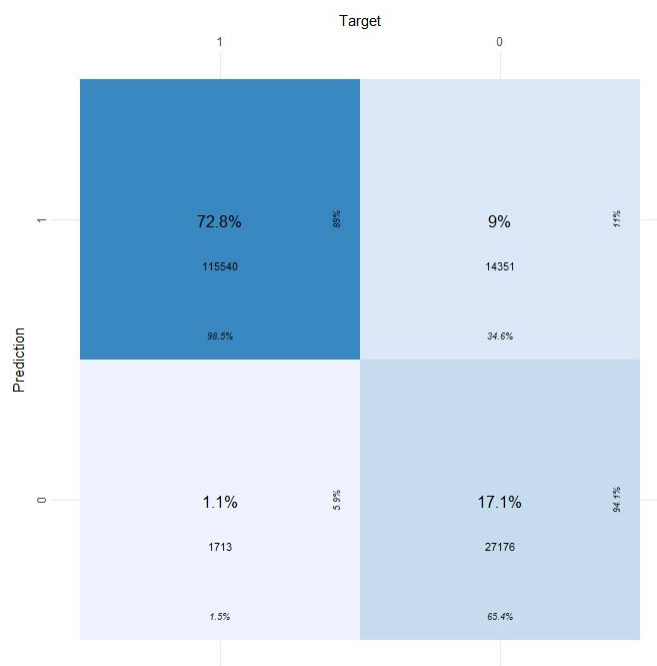


FIGURE 48 – Matrice de confusion du Random Forest pour prédire la survenance d'un arrêt de travail

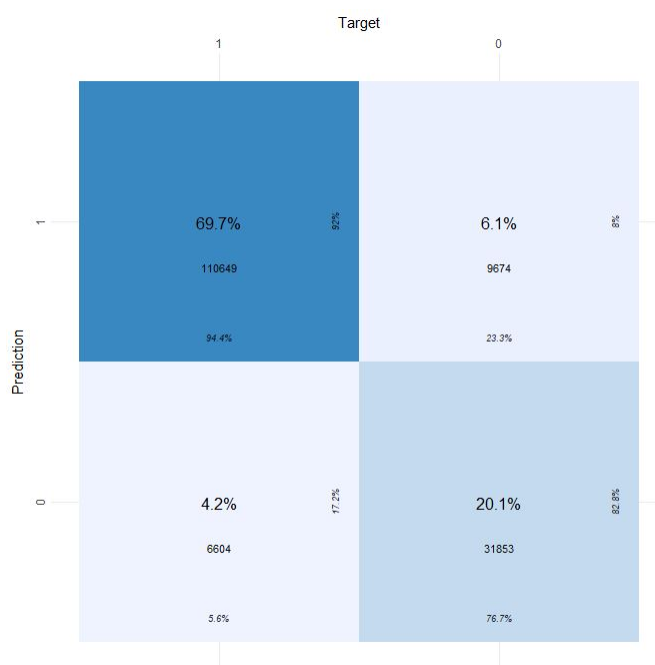


FIGURE 49 – Matrice de confusion du XGBoost pour prédire la survenance d'un arrêt de travail

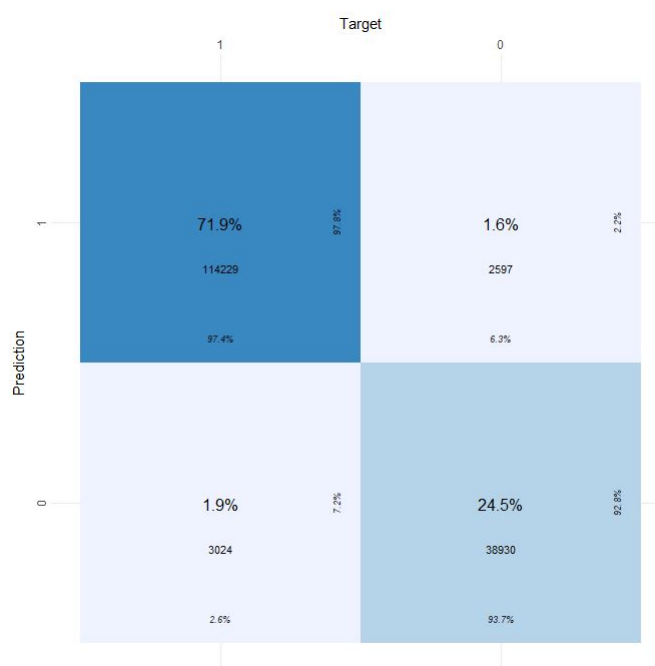


FIGURE 50 – Matrice de confusion Régression Logistique + Random Forest pour prédire la survenance d'un arrêt de travail

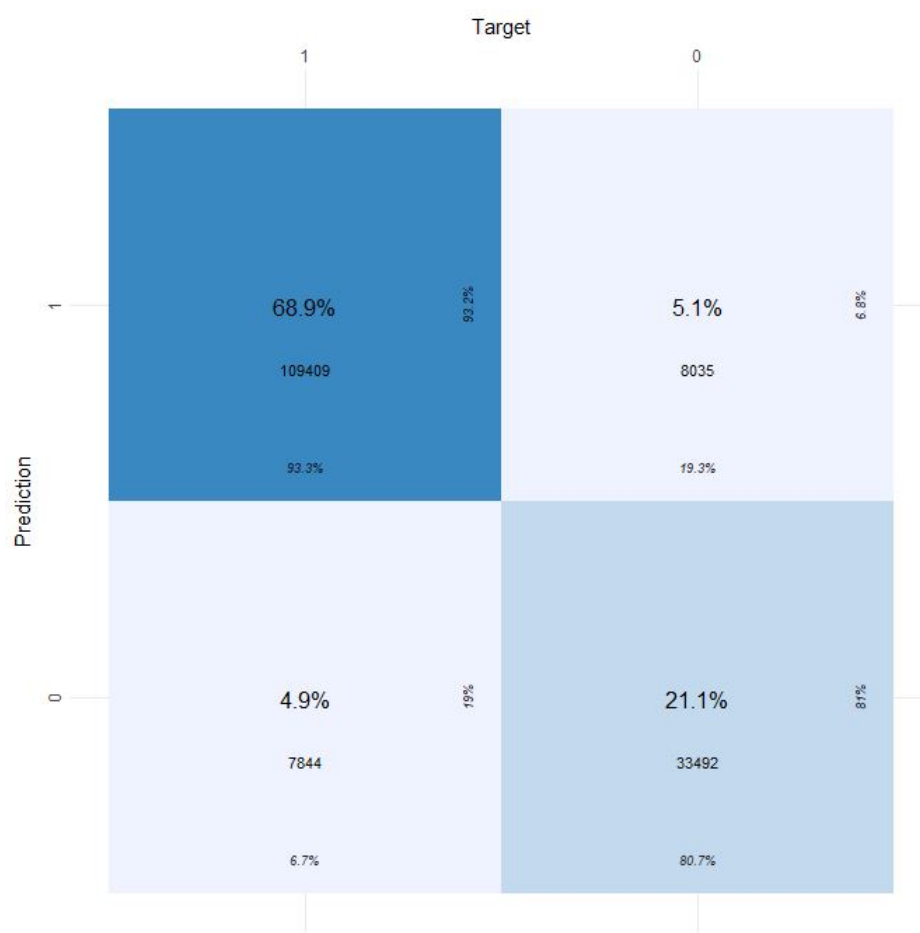


FIGURE 51 – Matrice de confusion Régression Logistique + XGBoost pour prédire la survenance d'un arrêt de travail