duas
Diplôme Universitaire des Actuaires de Strasbourg

UFR Mathématique Informatique Strasbourg

Mémoire présenté devant

l'UFR de Mathématique et Informatique

pour l'obtention du Diplôme Universitaire d'Actuaire de Strasbourg

et l'admission à l'Institut des Actuaires

le 10/12/2021

Par : Laure Weill

Titre: Amélioration des normes tarifaires de la Prévoyance grâce aux données DSN

Confidentialité : ☐ NON ☐ OUI Durée : ☐ 1 an ☒ 2 ans ☐ 3 ans ☐ 4 ans ☐ 5 ans

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

Signature :

Membres du jury de l'Unistra :

P. ARTZNER
J. BERARD
A. COUSIN
K.-T. EISELE
M. MAUMY-BERTRAND

Entreprise : AG2R La Mondiale

Directeur de mémoire en entreprise:
Nom : SAID I.
Signature : Imen SAID
Invité : SAID Imen

Nom : Imen SAID
Signature : SAID Imen

Jury de l'Institut des Actuaires : A.YOU et E. POYET

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

SAID Imen

Signature du candidat

.

Secrétariat : Mme Stéphanie Richard

Bibliothèque : Mme Christine Disdier

Résumé

Les acteurs de la Prévoyance collective font face aujourd'hui à des défis importants : le marché est de plus en plus concurrentiel et la dérive de l'arrêt de travail - déjà constatée avant la crise sanitaire - a été accélérée par la pandémie du COVID. Assureurs et institutions de Prévoyance (IP) doivent donc innover et saisir toutes les opportunités disponibles pour relever ces défis et mieux comprendre et maitriser le risque de l'arrêt de travail (AT).

La DSN (Déclaration Sociale Nominative) présente justement une opportunité forte pour les IP disposant d'un portefeuille Prévoyance sans affiliation (le cas de l'organisme d'accueil). Obligatoire depuis 2017 pour toutes les entreprises du secteur privé employant des salariés, la DSN s'appuie sur des données de la paie et permet aux organismes assureurs de la Prévoyance d'avoir une vision fine et actualisée de la réalisation de l'arrêt de travail. Ces données permettent donc de supprimer la censure dans les observations : une vraie révolution pour un actuaire habitué à l'asymétrie de l'information ! Sans la DSN, l'IP n'observe que la population sinistrée et uniquement les sinistres dépassant la franchise du contrat. Avec la DSN, l'IP peut observer toute la population assurée et tous les arrêts. Il devient donc possible de faire des travaux de segmentation permettant de caractériser la population à risque.
En addition à la disponibilité de la donnée, les acteurs du marché disposent de plus en plus d'outils de traitement de données massives (big data). Et, la communauté scientifique est bien avancée sur des algorithmes de segmentation basés sur la théorie du Machine Learning (ML). Ces algorithmes permettent d'expliciter des patterns dans des données volumineuses : une deuxième opportunité pour challenger les méthodes classiques de calcul du coût du risque et de sa tarification.

Dans ce mémoire, on se propose d'exploiter la DSN afin de challenger les normes tarifaires existantes sur le risque AT en le modélisant avec les méthodes du ML. On se focalise sur les arrêts de travail de courte durée : arrêts non observés par l'assureur.

En premier lieu, on présente les statistiques descriptives et une étude des corrélations entre les variables de la DSN et la réalisation de l'AT.

Ensuite, l'objectif sera de modéliser la loi de fréquence et la loi de maintien des arrêts qui durent moins de 6 mois sur certaines entreprises relevant des accords de branches « aide à domicile » et « propreté ». Pour ce faire, on teste et compare les performances de différents algorithmes : CART, GLM, Random Forest et XGBoost.

Finalement, les deux modèles sont agrégés pour avoir un coût total (fréquence x coût). Celui-ci permettra par la suite de tarifier les arrêts pour l'entreprise assurée.

Mots clés : Prévoyance collective, déclaration sociale nominative, modélisation, incapacité, arrêt de travail, fréquence, durée de maintien en incapacité, Kaplan Meier, Forêts aléatoires, GAM, GLM, CART, XGBoost, normes tarifaires

Summary

Group provident insures are now facing major challenges: the market is increasingly competitive and the slippage of the cost of sick leave - already observed during the recent years - has been accelerated by the COVID pandemic. Insurers and provident institutions (PI) must therefore innovate and seize all available opportunities to meet these challenges and better understand and control the risk of sick leave (AT).

The DSN (Nominative Social Declaration) presents a strong opportunity for PIs with a non-affiliation provident portfolio (the case of the host organization). Mandatory since 2017 for all companies in the private sector with employees, the DSN is based on payroll data and allows provident insurance organizations to have a detailed and up-to-date information about the sick leave cost. These data eliminate censorship in observations: a real revolution for an actuary accustomed to information asymmetry! Without the DSN, the PI only observes the affected population and only claims exceeding the policy deductible. With the DSN, the PI can observe the entire insured population and all the claims. This comprehensive information makes it possible to carry on modelling and segmentation work to characterize population with high risk of sick leave and its probable duration.
In addition to the availability of data, insurers have more and more tools for processing big data. The scientific community is well advanced on segmentation algorithms based on Machine Learning (ML) theory. These algorithms make it possible to explicit patterns in big data: a second opportunity to challenge traditional methods of calculating the cost of risk and its pricing.

The purpose of this thesis is to use the DSN to challenge the existing pricing standards on work stoppage risk by modelling it with ML methods. It focuses on short-term work stoppages: those that are not observed by the insurer without the DSN.

First, descriptive statistics and correlations between the variables of the DSN and the stoppage risk are presented.

Then, the objective will be to model frequency and duration of sick leaves that last less than 6 months on certain companies falling under the "home help" and "cleaning" branch agreements. To do this, we test and compare the performance of different algorithms: CART, GLM, Random Forest and XGBoost.

Finally, we aggregate outputs of both models to have a total cost model (frequency x duration). This will then make it possible to price the risk.

Keywords: collective provident insurance, nominative social declaration, modelling, incapacity, work stoppage, frequency, duration of incapacity, Kaplan Meier, Random forests, GAM, GLM, CART, XGBoost, tariff standards

Note de synthèse

Plus que jamais, et dans le contexte de la crise sanitaire, la dérive de l'arrêt de travail constatée depuis quelques années devient visible par les différents acteurs. Au-delà d'être le signal d'un fort mal-être au travail, pour reprendre les termes de Fabien Piazzon[1], cette dérive génère un coût important pour la collectivité, les entreprises et les assureurs.

Dans ce mémoire, on se propose justement de challenger les méthodes classiques d'estimation du coût du risque de l'arrêt de travail en s'appuyant sur les données issues de la Déclaration Sociale Nominative (DSN) d'un côté et les méthodes de modélisation basées sur les techniques du Machine Learning (ML) de l'autre.

Avant d'accéder aux DSN, les données dont dispose l'assureur d'un contrat de Prévoyance collective sans affiliation ne concernent que les salariés sinistrés ainsi que leurs arrêts ayant une durée supérieure à la durée de franchise contractuelle. La tarification des arrêts de travail était donc établie à partir d'informations censurées. L'obligation des messages DSN depuis 2017 a permis aux organismes complémentaires de réduire cette censure et d'avoir - à chaque envoi mensuel - accès à l'information détaillée de toute la population assurée (sinistrée ou non) et à tous les sinistres (dépassant ou non la franchise contractuelle). Le risque est donc observé dans sa totalité et les techniques de segmentation individuelle deviennent adéquats pour modéliser ce risque et mieux estimer son coût.

La problématique du mémoire se situe dans le challenge des normes tarifaires des arrêts de travail grâce aux données de la DSN. Pour cela, les travaux principaux effectués tournent autour de la modélisation de l'arrêt de travail dans sa durée et sa fréquence grâce à des méthodes performantes de Machine Learning. Ainsi, une tarification peut être déduite.

Il est présenté ici les différentes étapes des travaux qui permettent d'aboutir à notre objectif.

1. Statistiques descriptives

Les statistiques descriptives permettent de caractériser la population observée et de cibler les variables qui ont - a priori - un effet sur la fréquence d'arrêt et sa durée :

- Le sexe ;
- Le salaire ;
- L'âge ;
- La distance du domicile du salarié à l'entreprise ;
- Le motif d'arrêt de travail ;
- La région ;
- La taille de l'établissement de travail ;
- La CSP ;
- Le mois d'arrêt ;
- Le secteur lourd, moyen ou léger de l'entreprise ;

[1] Directeur Délégué Vie et Bien Être au travail - Ayming

- La modalité de l'exercice (temps plein ou partiel).

Avant cette analyse exploratoire, et comme pour chaque exercice de modélisation, un travail sur les données est effectué :

2. Retraitements

La base de données subit plusieurs retraitements comme la recatégorisation de certaines variables, l'ajout de certaines variables, le traitement des valeurs manquantes, le dédoublonnage des arrêts et enfin la fusion des rechutes des arrêts.

Deux bases sont construites pour nos modélisations :

- Une base d'apprentissage à la maille salarié pour modéliser la fréquence d'arrêt.
- Une base d'apprentissage à la maille arrêt pour modéliser la durée de l'arrêt.

Les modélisations sont effectuées sur ces tables retraitées.

3. Modélisations

Une étude sur la modélisation des arrêts de travail est établie. Les lois de fréquence et de maintien de l'arrêt sont modélisées grâce à plusieurs algorithmes de Machine Learning : CART, Random Forest, XGBoost et GLM.

Ces modélisations sont construites sur les arrêts courts de moins de 6 mois car c'est sur ces arrêts que nous avions un manque de visibilité sans les données de la DSN. Elles sont réalisées sur un périmètre bien spécifique. Nous choisissons d'effectuer notre étude sur les branches « propreté » et « aide à domicile » qui représentent un enjeu important pour l'entreprise. En effet, elles contiennent un volume d'arrêts et un chiffre d'affaire conséquents. De plus, uniquement les années 2018 et 2019 sont utilisées pour l'étude, puisque la DSN est fiable à partir de 2018 seulement, et le Covid qui arrive en 2020 peut biaiser nos modélisations.

Les modélisations sont développées en plusieurs étapes.

1. **Les lois de fréquence et de durée sont modélisées avec les algorithmes CART, Random Forest et XGBoost.**

2. **Les modèles CART, Random Forest et XGBoost sont ensuite comparés.**

La comparaison se fait à l'aide des indicateurs suivants : Gini, RMSE, MSE et MAE. Les courbes de Lorenz et les courbes Lift sont aussi analysées.

Cette étude est illustrée sur le modèle de fréquence.

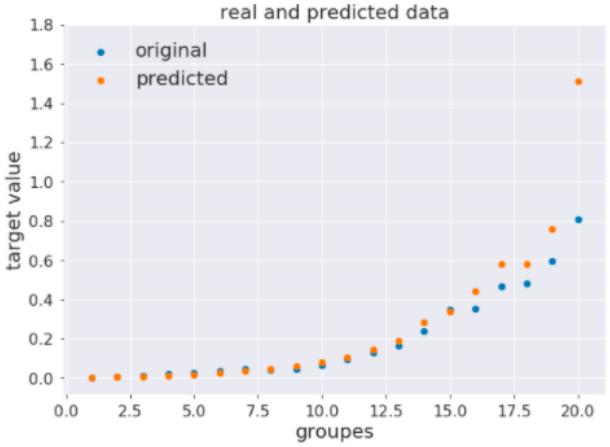
➢ <u>Comparaison des indicateurs de performance globaux</u>

	CART	Random Forest	XGB
RMSE	6,88	6,55	6,46
MAE	1,76	1,73	1,58
MSE	43,40	42,94	41,82
Gini	22%	26%	28%

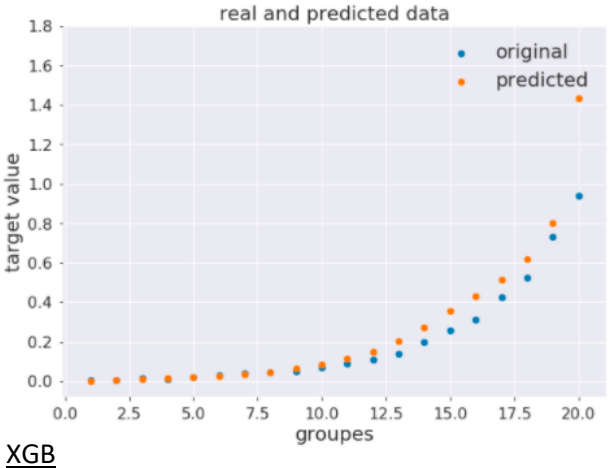
Les indicateurs RMSE et MAE qui mesurent l'erreur du modèle sont moins importants pour le XGB que pour les autres modèles. Pour le Gini qui mesure la qualité de segmentation du portefeuille, c'est l'inverse. Le XGBoost est donc sélectionné comme le meilleur modèle.

- Comparaison des courbes Lift

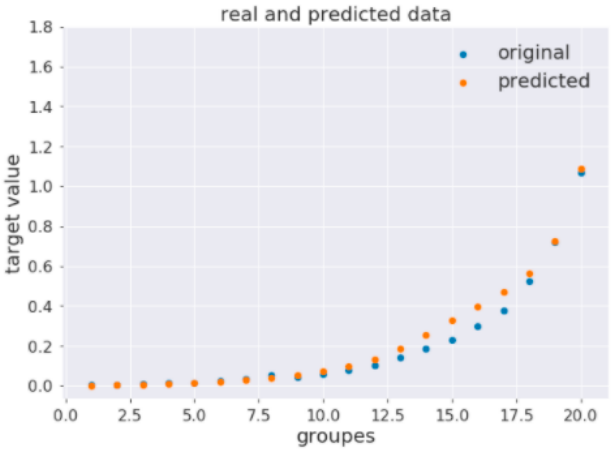
CART



Random Forest



XGB

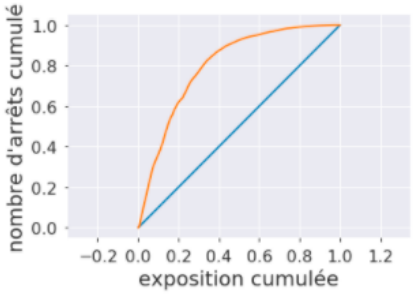


La courbe Lift est construite en séparant la population triée par niveaux de prédictions croissants en 20 échantillons de taille égale. Puis, pour chaque échantillon, deux valeurs sont calculées : la moyenne des prédictions et la moyenne des observations.

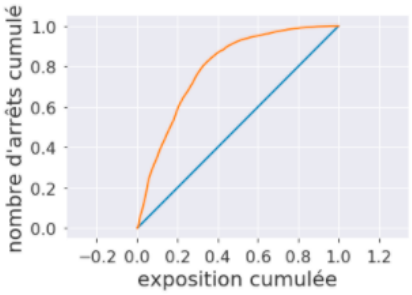
Les courbes Lift nous montrent donc que lorsque nous passons du CART, au Random Forest puis au XGBoost, la différence entre la moyenne des prédictions et la moyenne des observations devient de moins en moins importante, surtout pour les valeurs élevées. Cela confirme que le modèle XGB est le plus performant, surtout pour les fréquences importantes : il reste donc meilleur en moyenne et en valeurs extrêmes.

➢ Comparaison des courbes de Lorenz

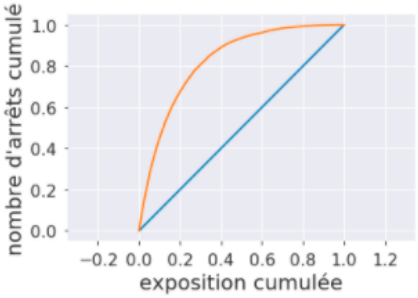
CART



Random Forest



XGB



La courbe de Lorenz est construite en triant les observations et les expositions selon les prédictions décroissantes et en traçant les expositions cumulées observées et normées en fonction des prédictions cumulées triées normées.

En analysant les courbes de Lorenz, nous constatons qu'elle est de plus en plus concave (en allant du CART au Random Forest au XGBoost) et que la courbe croît plus vite. Le XGBoost améliore donc la capacité à bien classifier les fréquences.

Pour le modèle de durée, la comparaison se base sur le même raisonnement et nous permet de sélectionner aussi le XGBoost comme modèle le plus performant.

Après avoir comparé ces trois méthodes de Machine Learning, des modélisations GLM sur la fréquence et la durée de l'arrêt sont introduites.

3. Les lois de fréquence et de durée sont modélisées avec le modèle GLM.

Ces modèles sont traités à part car ils sont construits avec un outil différent de Python (utilisé pour nos algorithmes CART, Random Forest et XGBoost). C'est un outil « clé en main » qui permet de prendre en compte plus de variables pour les modélisations.

Les modèles GLM sont comparés avec les modèles XGBoost qui donnent les meilleures modélisations parmi les trois précédentes

4. Les modèles XGBoost et GLM sont comparés à l'aide d'indicateurs de performances.

La comparaison est illustrée sur le modèle de fréquence.

	XGB	GLM
RMSE	6,46	2,22
MAE	1,58	0,71
MSE	41,82	4,93
Gini	28%	47,33%

Les indicateurs d'erreur RMSE, MAE et MSE sont plus faibles pour le modèle GLM que pour le XGB et le Gini est plus important pour le GLM. Nous sélectionnons donc le modèle GLM comme le plus performant des modèles implémentés dans ce mémoire.

Après sélection des modèles GLM pour les lois de fréquence et de durée, les deux lois sont agrégées pour obtenir la loi finale utilisée pour la tarification.

4. Agrégation des lois de fréquence et de durée par une multiplication

Une fois que les lois de fréquence et de maintien en arrêt sont construites, elles peuvent être agrégées par une multiplication. Nous obtenons ainsi le modèle prédisant la durée totale des arrêts sur un an pour un salarié affecté à un contrat de Prévoyance collective. C'est cette loi résultante qui est utilisée pour créer nos normes tarifaires.

5. Tarification

Pour la tarification, les variables utilisées sont celles qui sont les plus importantes pour les modèles de fréquence et de durée. Le tarif est donc basé sur les critères suivants :

- APE ;
- Montant de rémunération annuel moyen ;
- Effectif de l'entreprise ;
- Ancienneté moyenne ;
- Commune où se situe l'entreprise ;
- Age moyen ;
- Rémunération totale de l'entreprise.

Le commercial doit renseigner les données démographiques énoncées ci-dessus pour obtenir le tarif correspondant au profil entreprise.

C'est là qu'intervient notre modélisation GLM du nombre de jours total d'arrêts sur un an. Pour obtenir ces prédictions, le modèle GLM a estimé les paramètres β de la relation ci-dessous par la méthode du maximum de vraisemblance.

$$\hat{Y} = \mathbb{E}\,[Y \mid X = x] = g^{-1}\left(\sum_{\substack{i\epsilon[\![1,p]\!] \\ j\in[\![1,q_i]\!]}} x_{i,j} * \hat{\beta}_{i,j}\right) = \exp\left(\sum_{\substack{i\epsilon[\![1,p]\!] \\ j\in[\![1,q_i]\!]}} x_{i,j} * \hat{\beta}_{i,j}\right)$$

Avec :

- $\hat{Y}$: la prédiction du modèle ;
- X : le vecteur des variables $X_{i\epsilon[\![1,p]\!]}$ et x son observation ;
- g : la fonction de lien log ;
- p : le nombre de variables ;
- q_i : le nombre de modalités associé à la variable X_i ;
- $x_{i,j}$: variable n°i pour sa modalité n°j encodée en binaire : 1 si X contient cette réalisation, 0 sinon ;
- $\hat{\beta}_{i,j}$: le coefficient estimé pour la modalité n°j de la variable X_i.

Voici comment obtenir une prédiction du nombre de jours total d'arrêt(s) sur un an en appliquant la formule ci-dessus :

$$\text{Prédiction} = e^{intercept} \times e^{\beta_{l,m}} \times \ldots \times e^{\beta_{n,o}} \times \ldots$$

$$\text{Prédiction} = Offset \times Coefficient_{age} \times \ldots \times Coefficient_{salaire} \times \ldots$$

Ces coefficients représentent donc la correction qui est appliquée à l'offset qui est le nombre de jours d'arrêts total moyen prédit sans aucuns effets de variables. Ils permettent d'adapter cette moyenne à un profil client. Ce sont donc ces coefficients qui sont nécessaires à notre tarification.

En renseignant les données démographiques du profil client entreprise, les coefficients correcteurs du tarif s'adaptent et le tarif est obtenu en fonction du niveau de garantie.

La formule utilisée pour calculer la prime commerciale est la suivante :

$$\text{Prime commerciale annuelle} = \frac{\text{Prime pure}}{1 - \text{Chargements}}$$

Où Prime pure = (Niveau de garantie × Probabilité d'arrêt × $\text{Coefficient}_{\text{APE}}$ × $\text{Coefficient}_{\text{MONTANT REMUNERATION ANNUEL MOYEN}}$ × $\text{Coefficient}_{\text{EFFECTIF}}$ × $\text{Coefficient}_{\text{ANCIENNETE MOYENNE}}$ × $\text{Coefficient}_{\text{COMMUNE}}$ × $\text{Coefficient}_{\text{AGE MOYEN}}$ × $\text{Coefficient}_{\text{REMUNERATION TOTALE ENTREPRISE}}$ - Niveau remboursement Sécurité sociale - Niveau remboursement employeur) × Rémunération moyenne

La rémunération moyenne est la rémunération totale de l'entreprise sur un an divisée par l'effectif.
La probabilité d'arrêt sur un an est équivalente à l'offset du modèle divisé par 365.

En conclusion, nous sommes parvenus à challenger la tarification en mettant en place de nouvelles modélisations utilisant le Machine Learning. La performance du modèle a donc été poussée au maximum en sélectionnant le GLM.
De plus, la DSN nous a permis d'utiliser des données volumineuses et très fines au niveau du salarié, ce qui a contribué à la précision de prédiction des modèles. La connaissance nouvelle des non-sinistrés dans notre portefeuille et des arrêts dont la durée est inférieure à la franchise permet d'aboutir à des modélisations plus justes de la fréquence et la durée d'arrêt. Cela améliore ainsi la tarification pour les arrêts courts.
Mais il est possible d'améliorer encore les modélisations, surtout le modèle de durée. En effet, celui-ci présente une grande dispersion dans les valeurs qui fait diminuer la performance de prédiction du modèle. Une nouvelle piste s'ouvre donc et il serait envisageable de modifier la variable numérique de la durée d'un arrêt en une variable catégorielle avec des tranches de durées. Ainsi, le modèle pourrait être plus précis dans sa prédiction.
Enfin, nos modélisations ont mis en évidence des variables jugées comme les plus importantes pour la tarification. Celles-ci ne sont pas forcément les mêmes que dans le tarificateur précédent. Les critères de tarification sont donc renouvelés. Cependant, il serait intéressant de forcer des variables à être utilisées dans la tarification comme le sexe et la CSP qui sont couramment pris en compte pour un tarif. C'est là qu'intervient le besoin métier et nous devrions peut-être surpasser la théorie de nos modèles pour y répondre.
Pour aller encore plus loin dans le projet, une étude sur les arrêts de plus de 6 mois pourrait être envisagée afin de comparer la modélisation des arrêts longs avec et sans la prise en compte des données DSN.

Synthesis note

More than ever, and in the context of the health crisis, the drift of sick leave cost observed in recent years is becoming even more indeniable. Beyond being the signal of a strong malaise at work, to use the words of Fabien Piazzon[2], this slippage generates a significant cost for the community, businesses and insurers.

In this thesis, we aim to challenge the classic methods of estimating the cost of sick leave by relying on the data from the Nominative Social Declaration (DSN) on the one hand and the methods of modelling based on Machine Learning (ML) techniques on the other.

Before accessing the DSN, the data available to the insurer of a collective provident insurance contract without affiliation only concerns the affected employees as well as their stoppages lasting longer than the contractual deductible. The pricing of work stoppages was therefore established on the basis of censored information. The obligation of DSN since 2017 has enabled complementary organizations to reduce this censorship and to access detailed information of the entire insured population (affected or not) and to all claims (exceeding or not the contractual deductible). The risk is therefore observed in its entirety and individual segmentation techniques become adequate to model this risk and better estimate its cost.

The problem of the thesis lies in the challenge of pricing standards for work stoppages thanks to the data of the DSN. To do so, we start by modelling frequency and duration of work stoppage using high-performance Machine Learning methods. Then, a premium can be deducted.

The various stages of the work that lead to our goal are presented here.

1. Descriptive statistics

Descriptive statistics enables us to characterize the observed population and to target the variables which - a priori - have an effect on the frequency of quitting and its duration:

- Gender ;
- Salary ;
- Age;
- Distance from the employee's home to the company ;
- Reason for stopping work ;
- Region ;
- Size of the company ;
- CSP ;
- Month of work stoppage ;
- Heavy, medium or light sector of the company ;
- Modality of the exercise (full or part-time).

Before this exploratory analysis, and as for each modelling, reprocessing on the data is carried out:

[2] Assistant director Life and Wellbeing at Work - Ayming

2. Reprocessing

The database is subject to several reprocessing such as the re-categorization of certain variables, the addition of certain variables, the treatment of missing values, the deduplication of work stoppages and finally the merging of relapses from work stoppages.

Two databases are built for our models:

- A learning database with one line per employee, to model the stop frequency.

- A learning database with one line per work stoppage, to model the duration of the stoppage.

The modelling is carried out on these reprocessed tables.

3. Modelling

A study on the modelling of work stoppages is established. The laws of frequency and duration are modelled using several Machine Learning algorithms: CART, Random Forest, XGBoost and GLM.

These models are built on short work stoppages of less than 6 months because it is on these that we had a lack of visibility without the data from the DSN. They are carried out on a very specific perimeter. We have chosen to perform our study on the “cleaning” and “home help” branches, which represent an important part for the company portfolio. Indeed, they contain an important volume of work stoppages and a sizable turnover. In addition, just the years 2018 and 2019 are used for the study, since the DSN is reliable from 2018 only, and Covid coming in 2020 can bias our models.

The models are developed in several stages.

1. The frequency and duration laws are modelled with the CART, Random Forest and XGBoost algorithms.

2. Then, the CART, Random Forest and XGBoost models are compared.

The comparison is made using the following indicators: Gini, RMSE and MAE. Lorenz curves and Lift curves are also analysed.

This study is illustrated on the frequency model.

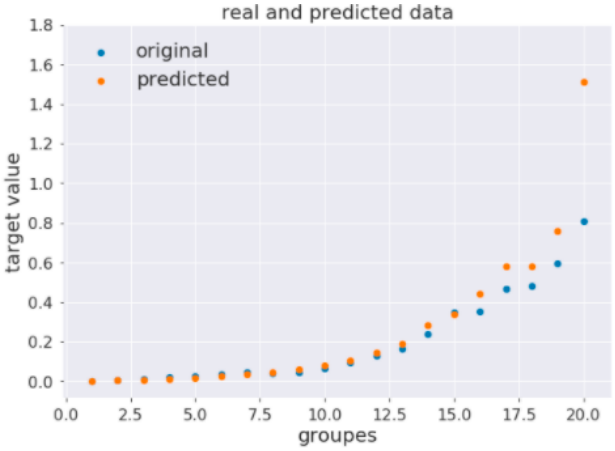
- Comparison of global performance indicators

	CART	Random Forest	XGB
RMSE	6,88	6,55	6,46
MAE	1,76	1,73	1,58
MSE	43,40	42,94	41,82
Gini	22%	26%	28%

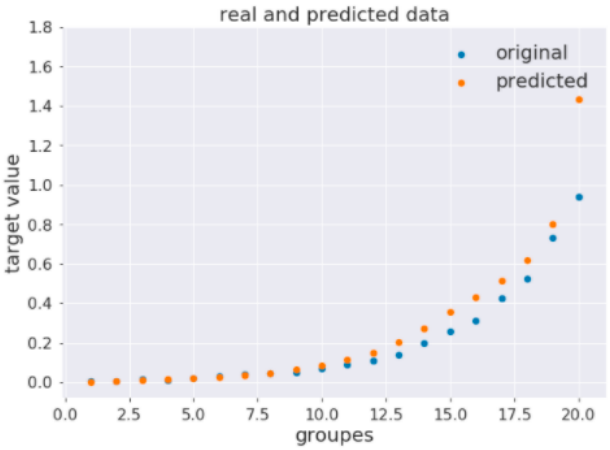
The RMSE and MAE indicators that measure model error are less important for the XGB than for other models. For the Gini, which measures the quality of portfolio segmentation, it’s the opposite. We therefore select the XGBoost as the best model.

➢ Comparison of Lift curves

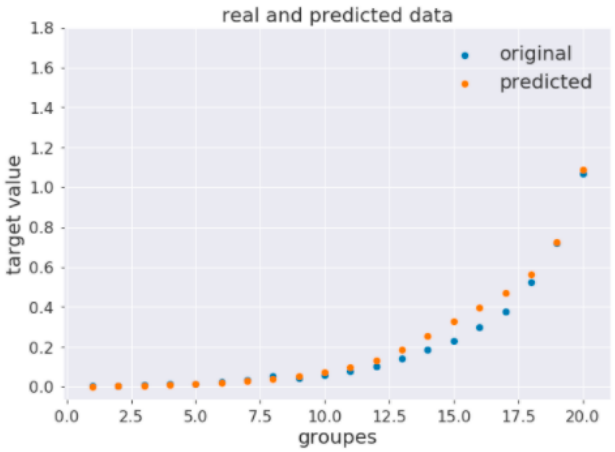
CART



Random Forest



XGB

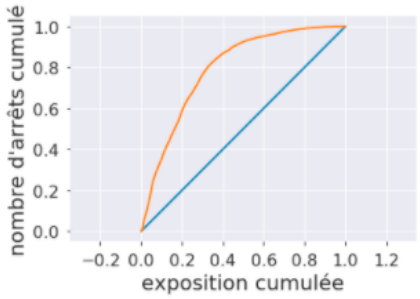


The Lift curve is built by separating the population sorted by increasing prediction levels into 20 samples of equal size. Then, for each sample, two values are calculated: the average of the predictions and the average of the observations.

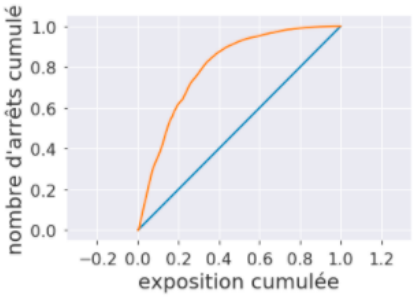
The Lift curves therefore shows us that when we look at CART, Random Forest and then XGBoost, the difference between the average of the predictions and the average of the observations becomes less and less important, especially for high values. This confirms that the XGB model is the most efficient, especially for high frequencies.

➢ Comparison of Lorenz curves

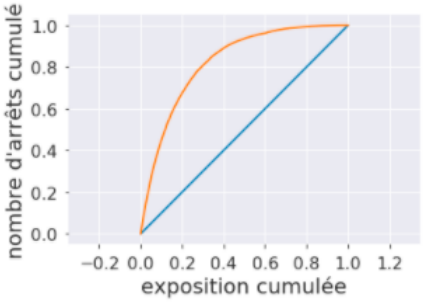
CART



Random Forest



XGB



The Lorenz curve is built by sorting observations and exposures according to decreasing predictions and by plotting the normalized cumulative observed exposures according to the sorted normalized cumulative predictions.

By analysing the Lorenz curves, we see that it is more and more concave (going from CART to Random Forest to XGBoost) and that the curve is increasing faster. XGBoost therefore improves the ability to properly classify frequencies.

For the duration model, the comparison is based on the same indicators and allows us to also select the XGBoost as the most efficient model.

After comparing these three Machine Learning methods, GLM models on the frequency and duration of the work stoppage are introduced.

1. **The frequency and duration laws are modelled with the GLM model.**

These models are treated separately because they are built with a different tool from Python (used for our CART, Random Forest and XGBoost algorithms). It is a turnkey solution which allows more variables to be considered for modelling.

The GLM models are compared with the XGBoost models which yield the best predictions among the three previous ones.

2. **XGBoost and GLM models are compared using performance indicators.**

We illustrate the comparison on the frequency model.

	XGB	GLM
RMSE	6,46	2,22
MAE	1,58	0,71
MSE	41,82	4,93
Gini	28%	47,33%

The RMSE, MAE and MSE error indicators are lower for the GLM model than for the XGB and the Gini is larger for the GLM. We therefore select the GLM model as the most efficient of the models implemented in this thesis.

After selecting the GLM models for the frequency and duration, we aggregate the two laws to obtain the final law used for pricing.

4. Aggregation of the frequency and duration laws by a multiplication

Once the frequency and duration laws are built, we can aggregate them by multiplication. We thus obtain the model predicting the total duration of a work stoppage over one year for an employee assigned to a collective provident insurance policy. It is this resulting law that is used to create our tariff standards.

5. Pricing

For pricing, the variables used are the most important for the frequency and duration models. The price is therefore based on the following criteria:

- APE ;

- Average annual payment amount ;

- Headcount of the company ;

- Average seniority in the firm ;

- Municipality where the company is located ;

- Average age ;

- Total remuneration of the company.

The commercial must fill in the demographic data set out above to obtain the price corresponding to the company profile.

This is where our GLM modelling of the total number of work stoppage(s) in a year is introduced. To obtain these predictions, the GLM model estimated the parameters β of the relation below by the maximum likelihood method.

$$\hat{Y} = \mathbb{E}[Y \mid X = x] = g^{-1}\left(\sum_{\substack{i\in[\![1,p]\!] \\ j\in[\![1,q_i]\!]}} x_{i,j} * \hat{\beta}_{i,j}\right) = \exp\left(\sum_{\substack{i\in[\![1,p]\!] \\ j\in[\![1,q_i]\!]}} x_{i,j} * \hat{\beta}_{i,j}\right)$$

With:

- $\hat{Y}$: model prediction ;
- X : vector of variables $X_{i\in[\![1,p]\!]}$ and x its observation ;
- g : link function log ;
- p : number of variables ;
- q_i : number of modalities associated with the variable X_i ;
- $x_{i,j}$: variable n°i for its modality n°j encoded in binary: 1 if X contains this observation, 0 otherwise ;
- $\hat{\beta}_{i,j}$: estimated coefficient for the modality n°j of the variable X_i.

Here's how to get a prediction of the total number of work stoppage(s) over a year by applying the above formula:

$$\text{Prediction} = e^{intercept} \times e^{\beta_{l,m}} \times \ldots \times e^{\beta_{n,o}} \times \ldots$$

$$\text{Prediction} = Offset \times Coefficient_{age} \times \ldots \times Coefficient_{salary} \times \ldots$$

These coefficients therefore represent the correction that is applied to the offset, which is the predicted average total number of days of work stoppage without any effect of variables. They allow this average to be adapted to a customer profile. It is therefore these coefficients that are necessary for our pricing.

By filling in the demographic information of the client company profile, the price is calculated according to the corresponding correction coefficients and the desired level of guarantee.

The formula used to calculate the commercial premium is as follows:

$$\text{Annual commercial premium} = \frac{\text{Pure premium}}{1 - \text{Expenses}}$$

Where $\text{Pure premium} = (\text{Guarantee level} \times \text{Work stoppage probability} \times \text{Coefficient}_{\text{APE}} \times \text{Coefficient}_{\text{AVERAGE ANNUAL REMUNERATION AMOUNT}} \times \text{Coefficient}_{\text{HEADCOUNT}} \times \text{Coefficient}_{\text{AVERAGE SENIORITY}} \times \text{Coefficient}_{\text{MUNICIPALITY}} \times \text{Coefficient}_{\text{AVERAGE AGE}} \times \text{Coefficient}_{\text{TOTAL REMUNERATION COMPANY}} - \text{Social Security reimbursement level} - \text{Employer reimbursement level}) \times \text{Average remuneration}$

Average remuneration is the company's total remuneration over one year divided by headcount.

Work stoppage probability over one year is equivalent to the model offset expressed in years (divided by 365).

In conclusion, we have succeeded in challenging the pricing by implementing new models using Machine Learning. We therefore pushed the performance of the model to the maximum by selecting the GLM.
Moreover, the DSN allowed us to use large and very detailed data with regards to employee, which contributed to the prediction accuracy of the models. New knowledge of the non-affected in our portfolio and work stoppages lasting less than the deductible makes it possible to achieve more accurate frequency and duration models. This improves the pricing for short sick leaves.
But it is possible to further improve the modelling, especially the duration model. Indeed, this one presents a large dispersion in the target variable which leads to a poor prediction performance. To increase the performance, it is possible to encode differently this variable into a categorical one with duration periods.
Finally, our models revealed variables considered to be the most important for pricing. These are not necessarily the same as in the previous pricing standard. The pricing criteria are therefore renewed. However, it would be interesting to force variables to be used in pricing such as gender and CSP which are commonly considered for a tariff. This is where the business rules must be considered and maybe we should go beyond the theory of our models to meet it.
To go even further in the project, a study on work stoppages of more than 6 months could be explored to compare the modelling of long work stoppages with and without taking DSN data into account.

Remerciements

Tout d'abord, je tiens à remercier ma tutrice de stage Imen SAID, qui m'a donné l'opportunité de travailler sur ce projet si riche et intéressant. J'aimerais aussi la remercier pour m'avoir apporté son soutien et accordé sa confiance dans les travaux que j'ai effectués, tout cela dans une atmosphère particulièrement agréable.

Je remercie aussi Anne-Charlotte DE RAIGNIAC, Michael JACCAZ, Auryane HOAREAU, Fabiola DEWULF, Abdelilah ZOUHADI, Jean FERRERO et Timothe KUEHM qui m'ont été d'une grande aide tout au long du stage, qui ont su enrichir mes connaissances et m'ont permis d'avancer dans le projet.

Merci aussi à toutes les personnes avec qui j'ai pu collaborer, sans lesquels je n'aurais pu parvenir à la réalisation de ce travail.

Je tiens également à remercier mon tuteur académique Jean BERARD pour ses précieux conseils.

Enfin, je remercie du fond du cœur ma famille et mes amis, et tout particulièrement Uriel qui m'a beaucoup soutenu et encouragé dans la réalisation de ce mémoire.

Liste des abréviations

- **ACM : Analyse en Composantes Multiples**
- **ACP : Analyse en Composantes Principales**
- **AE : Attestation Employeur**
- **AIC : Akaike Information Criterion**
- **APE : Activité Principale Exercée**
- **AT : Arrêt de Travail**
- **BCAC : Bureau Commun d'Assurances Collectives**
- **BIC : Bayesian Information Criterion**
- **CART : Classification And Regression Trees**
- **CCN : Convention Collective Nationale**
- **CCVRP : Caisse Nationale de Compensation des VRP (vendeur, représentant et placier)**
- **CDD : Contrat à Durée Déterminée**
- **CDI : Contrat à Durée Indéterminée**
- **DADS-U : Déclaration Annuelle des Données Sociales Unifiée**
- **DMMO : Déclaration de Mouvements de Main d'Œuvre**
- **DSN : Déclaration Sociale Nominative**
- **DVA : Données à Valeur Ajoutée**
- **DUCS : Déclaration Unifiée des Cotisations Sociales**
- **EMMO : Enquête sur les Mouvements de Main d'Œuvre**
- **GAM : Generalized additive Model**
- **GLM : Generalized linear models**
- **IJ : Indemnités Journalières**
- **IP : Institut de Prévoyance**
- **KM : Kaplan Meier**
- **MAE : Mean Absolute Error**
- **ML : Machine Learning**
- **MSE : Mean-Square-Error**
- **NEODeS : Norme d'Echange Optimisée des Données Sociales**
- **NIR : Numéro d'Inscription au Répertoire**
- **RF : Random Forest**
- **RH : Ressources Humaines**
- **RMSE : Root-Mean-Square-Error**
- **SIREN : Système d'Identification du Répertoire des ENtreprises**
- **SIRET : Système d'Identification du Répertoire des Établissements**
- **XGB : Extreme Gradient Boosting**

Table des matières

Introduction

Depuis 2017, la Déclaration Sociale Nominative (DSN) devient obligatoire pour les entreprises du secteur privé employant des salariés. Elle permet de centraliser les données employeurs provenant des ressources humaines via des déclarations mensuelles. Elle facilite ainsi la transmission des informations au sujet des salariés aux organismes qui en ont besoin. Cela simplifie aussi les calculs de cotisations.

Pour un assureur intervenant sur de la Prévoyance collective, l'intérêt de la DSN est double. Premièrement, elle apporte une connaissance de la population sous risque (sinistrés et non sinistrés). Secondement, elle fournit une connaissance de l'absentéisme, c'est-à-dire de l'ensemble des arrêts de travail quelque soient leurs durées. Ces deux apports viennent compléter les informations remontées dans les systèmes de gestion de Prévoyance collective qui recueillent seulement les salariés indemnisés post franchises.

L'assureur a alors la capacité d'être plus performant sur la connaissance du risque arrêt de travail de court terme pour des usages de tarification (qui est l'objectif de ce mémoire) ou pour du provisionnement. Il est essentiel de bien tarifer ce risque car il présente un enjeu important dans la couverture de Prévoyance collective des accords de branches, où souvent, les partenaires sociaux exigent la couverture des obligations de maintien de salaire, appelée usuellement par les assureurs « risque mensualisation ».

Plusieurs mémoires traitent de la modélisation des arrêts en passant par la construction de lois de fréquence et de maintien en incapacité. Les méthodes utilisées à ces fin rencontrées dans les travaux sont des méthodes statistiques non paramétriques comme la méthode de Kaplan Meier ou le modèle de Nelson Aalen, des modèles semi-paramétriques comme la régression de Cox ou encore des modèles paramétriques comme les régressions linéaires. Ces méthodes peuvent être retrouvées dans le mémoire de Caux J. « modélisation de l'entrée en incapacité temporaire de travail en Prévoyance collective ». Bien que ce dernier traite essentiellement des modélisations par les méthodes de Kaplan Meier ou Cox, il nous donne aussi une introduction aux modèles de Machine Learning avec l'algorithme CART.

Cependant, ces études ne prennent pas en compte les données de la DSN qui n'étaient pas fiables jusqu'à présent, et sont donc réalisées sur les bases de Prévoyance collective assez restreintes en informations sur les salariés. Par conséquent, les études ayant été effectuées sur l'arrêt de travail se contentent de modéliser les arrêts sur des salariés sinistrés uniquement et sur les arrêts dont la durée est supérieure à la durée de franchise.

Notre mémoire a donc pour objectif d'apporter une amélioration dans la modélisation de l'arrêt car notre étude porte sur les nouvelles données de la DSN qui sont plus nombreuses et plus fines. Nous nous référons ici au mémoire de Stefani L. « les méthodes de construction des barèmes tarifaires collectifs en arrêt de travail et apport des données de DSN » ayant prouvé l'apport des données DSN dans la modélisation de manière théorique. Cependant, dans celui-ci, ce ne sont pas des modèles de Machine Learning qui sont mis en place. Notre mémoire vient donc introduire ces nouvelles méthodes plus performantes appliquées aux bases de la DSN de manière pratique cette fois-ci.

L'objectif dans ce mémoire est donc d'améliorer les normes tarifaires des produits de Prévoyance collective sur l'arrêt de travail en utilisant les données de la DSN de 2018 et 2019. Pour challenger la tarification, des algorithmes de Machine Learning sont mis en place.

Tout d'abord, le contexte de l'absentéisme et les facteurs qui influencent l'arrêt au travail seront analysés. Puis, nous comprendrons comment fonctionne la DSN et les nombreux avantages que présente une étude sur ces données si riches plutôt qu'une étude sur les bases de la Prévoyance collective qui se limite au périmètre des sinistrés et des arrêts après franchise.

Ensuite, nous introduirons les statistiques descriptives et les études de corrélations qui nous permettront de cibler à priori les variables ayant le plus d'effets sur l'arrêt de travail. Mais aussi, les corrélations aideront à comprendre les différents liens qui existent entre les variables qui nous serviront à prédire les arrêts. Pour approfondir cela, des analyses en composantes principales et en composantes multiples seront mis en œuvre afin de déterminer les variables les plus importantes pour nos modèles. Ces analyses serviront aussi à éliminer de l'information qui serait redondante sur les modèles.

Enfin, nous aboutirons à la construction de deux lois pour les arrêts de moins de 6 mois : une qui prédit la fréquence d'arrêt pour un salarié dans un certain contrat de travail et une qui prédit la durée de l'arrêt. Elles seront élaborées à l'aide des méthodes de Machine Learning suivantes : GLM, CART, Random Forest et XGBoost. Par cette occasion, un outil de tarification externe basé sur des GLM sera évalué dans un objectif de défi pour l'IP.

Ces lois seront comparées grâce aux indicateurs de performance RMSE, MSE, MAE et Gini. Aussi, les courbes de Lorenz et les courbes Lift nous permettront d'évaluer la qualité des modèles dans leur détail. Finalement, les modèles qui seront jugés les plus performants vont être sélectionnés. Le modèle final prédisant la durée totale des arrêts sur un an sera alors obtenu par multiplication des lois de fréquence et de durée.

La prochaine étape sera la construction de normes tarifaires grâce à la loi finale qui nous permettra d'obtenir la probabilité d'arrêt qui sera multipliée par le montant de rémunération annuel moyen.

Partie 1 : contexte de l'étude

Cette partie sera introduite avec l'évolution de l'absentéisme et ses enjeux. Puis, nous nous intéresserons au coût de cet absentéisme (arrêt de travail) pour un organisme complémentaire et son estimation par les tables réglementaires du BCAC. Nous clôturerons en définissant la DSN, son évolution et son apport afin de comprendre l'enjeu du mémoire.

Section 1.1 : l'évolution de l'absentéisme et ses enjeux

« L'absentéisme s'est dégradé deux fois plus en un an de crise que la somme des trois années précédentes. La crise a révélé qu'il y avait urgence à agir contre cette tendance qui s'installe lourdement et durablement, afin d'éviter que cela ne se prolonge » [3], a exprimé Denis Blanc, directeur général du groupe Ayming, lors de l'exposition des résultats du 13ème baromètre de l'Engagement d'Ayming et AG2R La Mondiale.

Dans cette section, les multiples facteurs de l'absentéisme sont détaillés tout en présentant son évolution et ses causes. Nous mentionnons la crise du Covid 19 qui perturbe fortement les entreprises et leurs salariés en accélérant cette évolution. Enfin, les constats et les statistiques sur l'absentéisme au travail nous permettent de comprendre les différents enjeux de ce risque pour un assureur proposant des couvertures indemnisant les arrêts de travail tant pour le provisionnement que pour la tarification.

Beaucoup de facteurs favorisent l'augmentation du risque d'absentéisme au travail. Aux raisons de santé, peuvent s'ajouter la dégradation des conditions de travail, le niveau de stress, la qualité de vie et le vieillissement de la population. En effet, les chiffres de l'INSEE[4] nous montrent qu'en France, les taux d'activité pour les tranches de 15 à 64 ans augmentent d'années en années passant de 75,1% en 2014 à 75,7% en 2018 pour les hommes. Pour les femmes, le taux d'activité passe de 67,1% à 68,1%. Cette évolution est la conséquence de la réforme des retraites de 2014[5] qui tend entre autres à rallonger la période d'activité des séniors. Ce vieillissement pourrait avoir un impact sur l'absentéisme, les personnes âgées étant plus fragiles et donc plus sujettes aux arrêts de travail.

Les différentes raisons de l'augmentation des arrêts maladie ont fait l'objet d'une étude par Malakoff Humanis :

[3] https://www.gazettenormandie.fr/article/la-crise-a-accelere-l-absenteisme-et-le-desengagement-des-salaries
[4] https ://www.insee.fr/fr/statistiques/4501603?sommaire=4504425#tableau-figure2_radio1
[5] https://www.la-retraite-en-clair.fr/retraite-France monde/reformes-retraites-1993-2014/reforme-retraites-2014/ : « L'allongement de la durée d'assurance requise »

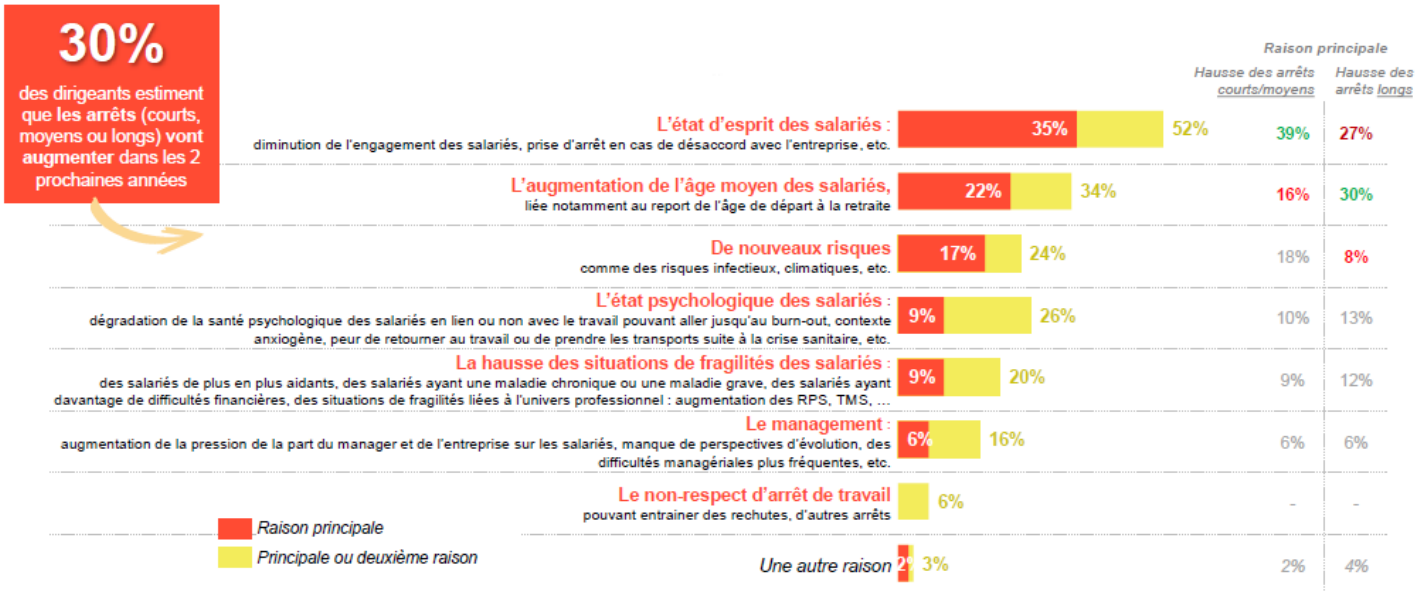


Figure 1 : raisons de l'augmentation des arrêts maladie[6]

En France, ces dernières années, une dégradation du taux d'absentéisme est observée. Entre 2017 et 2018, une hausse de 8% du taux d'absentéisme est constatée, soit un taux d'absentéisme moyen de 5,10% et 18, 6 jours d'absence en moyenne en 2018[7]. En 2019, le taux d'absentéisme est de 5,11% et 18,7 jours d'absence en moyenne. Les taux d'absentéisme peuvent être présentés plus précisément par secteur d'activité :

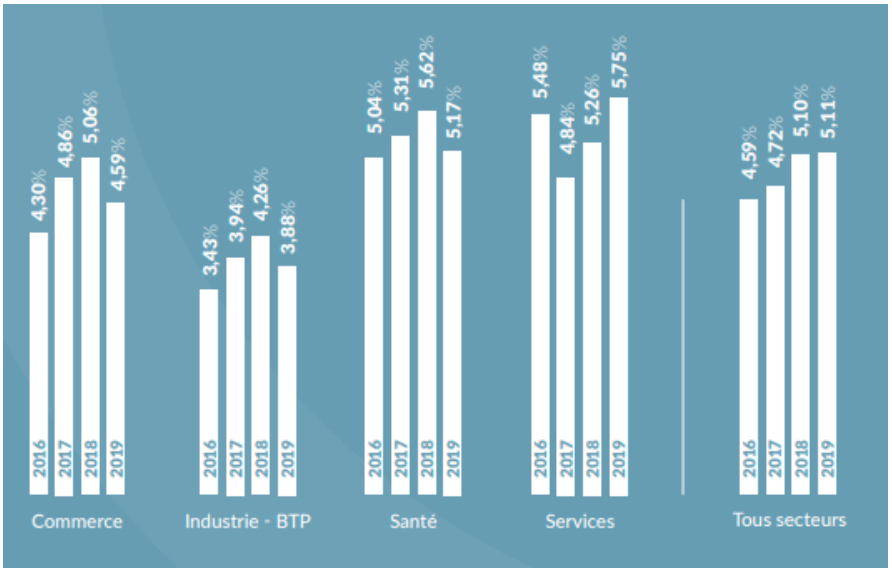


Figure 2 : évolution de l'absentéisme par secteur d'activité[8]

[6] Source : Baromètre annuel Absentéisme 2020 - Malakoff Humanis
[7] Etude Ayming, entreprises et salariés
[8] Etude Ayming, entreprises et salariés

Une étude de AG2R La mondiale et de l'Enquête Ayming-Kantar TNS a été réalisée en 2019 sur 45 403 entreprises privées ayant 1 921 000 employés, enquêtant sur les arrêts maladie et les accidents du travail professionnels. Il en ressort de cette étude plusieurs constats[9] :

- L'homogénéité de l'absentéisme parmi les différentes régions de France.
- Une baisse de l'absentéisme dans les domaines du commerce, de l'industrie et de la santé aux alentours de 9% et une hausse pour le secteur des services de 9%.
- Une évolution de l'absentéisme plus forte pour les salariés à temps partiel que pour les salariés à temps plein à raison de 1% de plus.
- L'absentéisme de longue durée chez les jeunes en forte augmentation, soit de 34% entre 2017 et 2019. Mais, avec un écart d'absentéisme qui se réduit entre les jeunes et les plus âgés qui dépassent la quarantaine bien que le taux d'absentéisme continue de croître avec l'âge.
- Concernant le sexe, la différence qui continue de se creuser en 2019 entre le taux d'absentéisme des hommes qui est de 3,59% et celui des femmes qui est de 5,84%.
- Un écart significatif entre le taux d'absentéisme chez les cadres qui est de 3,71% contre 6,45% chez les non-cadres.
- Un ralentissement dans l'augmentation du taux d'absentéisme chez les non-cadres avec un taux passant de 6,21% à 6,45% de 2018 à 2019, mais un nouvel élan d'absentéisme chez les cadres avec un taux passant de 3,10% à 3,71% de 2018 à 2019.
- En ce qui concerne les motifs d'absence, un taux d'absentéisme qui diffère en fonction du temps d'absence, comme illustré ci-dessous :

Motif	Arrêts de moins de 3 mois	Arrêts de plus de 3 mois
Maladie	51%	70%
Accident du travail/maladie professionnel	29%	17%
Situation familiale	8%	6%
Burn out	10%	5%
Démotivation	2%	2%

Figure 3 : les motifs d'arrêts par durée d'absence[10]

Après avoir analysé les différentes statistiques sur l'absentéisme au cours de ces dernières années, nous nous intéressons aux conséquences de ces absences sur les salariés. L'enquête montre qu'une bonne proportion de salariés sont absents plus d'une fois en 2019, avec un pourcentage de 41%. C'est pourquoi, une attention particulière doit être portée sur la première absence et sa cause. Une des causes, pas des moindres, à la répétition de l'arrêt peut être le mauvais accueil au retour du salarié dans l'entreprise. 20% de la population interrogée affirme avoir été victime de cela. Voici ci-dessous le résultat d'un sondage sur le renouvellement de l'absence au travail suite à une mauvaise réintégration :

[9] Les statistiques sont aussi disponibles sous forme de graphique en **annexe 1**. Ils sont tirés du Baromètre Ayming 2020 sur l'absentéisme.
[10] Etude Ayming, entreprises et salariés

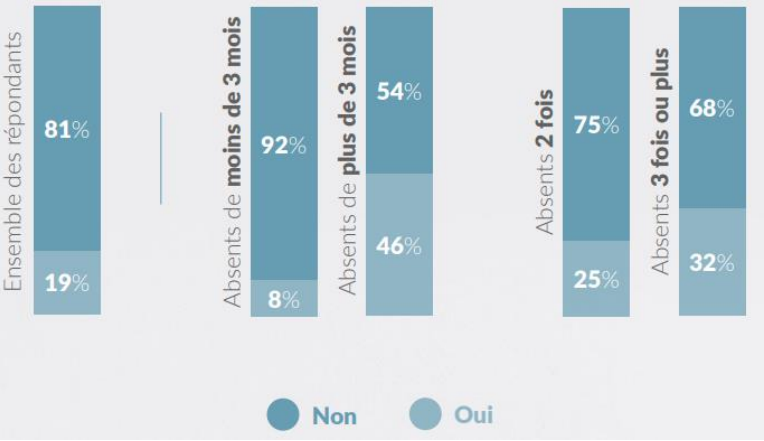


Figure 4 : sondage sur le renouvellement de l'absence au travail suite à une mauvaise réintégration[11]

L'étude montre aussi que l'influence des salariés sur ses collègues est notable. L'arrêt d'un salarié a des conséquences sur ces derniers qui sont eux-mêmes impactés. Dans les équipes où les absences sont fréquentes, deux fois plus de personnes sont absentes que dans les équipes ou il n'y a pas d'absences régulières. Cela peut s'expliquer par les impacts significatifs et multiples que ces arrêts ont sur les membres de l'équipe comme l'augmentation de la charge de travail, l'impact physique, la réorganisation du travail, la démotivation et la mauvaise ambiance au sein de l'équipe.
Il est donc primordial pour l'entreprise de mettre en action des plans pour lutter contre l'absentéisme. Effectivement, nous observons qu'en 2019, 16% des sondés déclarent avoir éviter l'absentéisme grâce aux plans d'actions mis en œuvre par l'entreprise. Les changements d'horaires, les aménagements de poste, les acceptations de congés, les entretiens avec les managers, les changements de poste et d'autres actions encore permettent de combattre les arrêts de travail.

Malgré ces efforts mis en place pour cette lutte contre ces absences, la crise du Covid 19 vient perturber fortement le monde du travail. Depuis l'année 2020, il est constaté de manière flagrante un absentéisme au travail sans précédent. D'après des études de Malakoff-Humanis, 26% des arrêts de travail en avril sont provoqués par la crise sanitaire et 19% en juin. Les personnes contaminées ou les personnes ayant renoncé à des soins pendant la période de confinement en sont une des causes. Mais encore, les appréhensions du retour au travail pour des raisons de risque de contamination sur le lieu du travail ou dans les transports en commun entre autres ont eu beaucoup d'impact sur l'absentéisme.

Dans le cadre de cette pandémie, il est intéressant de s'interroger sur les effets du changement des conditions de travail pendant la crise Covid sur les arrêts des salariés. Nous connaissons les conséquences négatives de celle-ci sur les taux d'absentéisme, comme celles que nous avons pu énoncer précédemment. Mais, nous nous questionnons aussi sur les aspects positifs qu'elle pourrait apporter notamment avec le passage du présentiel au télétravail. 84% des salariés en télétravail préféreraient poursuivre le télétravail après le confinement[12]. Ceci est dû aux avantages que cela apporte comme la suppression des trajets, l'amélioration de la qualité de vie familiale grâce au gain de temps et de la présence au foyer, la diminution de la fatigue liée aux déplacements, le gain de concentration et la flexibilité accordée dans les horaires. Toutefois, des effets négatifs sur le télétravailleur viennent atténuer cette situation qui peut paraitre idéale pour le salarié. Ce

[11] Etude Ayming, entreprises et salariés
[12] Selon une étude de Malakoff Humanis

nouveau mode de travail peut jouer tant sur le plan physique que psychique comme une baisse de la pratique sportive et plus généralement une dégradation de l'hygiène de vie.

Après avoir détaillé tous ces constats sur le risque d'arrêt de travail, nous comprenons que ce risque touche de multiples acteurs. De manière globale, tous les secteurs d'activités sont concernés par cette évolution mais le secteur de la Santé est particulièrement touché car ces absences favorisent les risques physiques et psychologiques de la population de travailleurs. Les personnes travaillant dans le transport, le commerce et les services sont les plus visés par cette crise d'absentéisme. De manière plus ciblée, l'absentéisme de longue durée pose un problème de déséquilibre organisationnel et de déstructuration au travail, il a donc un impact fort tant sur les employeurs que sur les salariés et sur la vie de l'entreprise. Nous ne négligeons pas les enjeux économiques et financiers que cela provoque car ces arrêts demandent des indemnités importantes. C'est ici que rentre en jeux l'état (Sécurité Sociale), les compagnies d'assurances et les mutuelles mais aussi les acteurs privés comme les instituts de Prévoyance qui doivent faire face à cela en mettant en place des provisionnements pour ce risque d'absentéisme.

Le Baromètre Ayming permet de mettre en évidence un risque en augmentation de manière durable et lourde (citation de Denis Blanc) avec des facteurs qui viennent accélérer le risque ou l'augmenter. On se propose d'effectuer une étude individuelle pour mieux cerner ces facteurs afin de challenger les normes tarifaires des produits de l'arrêt de travail.

Section 1.2 : le coût de l'arrêt de travail pour un organisme complémentaire

Un organisme complémentaire a la charge d'estimer le coût d'un risque qui dérive. Il faut comprendre le risque et le mécanisme de l'indemnisation.

1.2.1 La Prévoyance

D'après la loi Evin évoquée le 31 décembre 1989[14] la Prévoyance rassemble « les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant 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 ». Les organismes pouvant gérer ces opérations sont décrits dans l'article 1 (**annexe 2**). En résumé, la Prévoyance permet de pallier les pertes de revenus lors d'arrêts de travail ou de décès et propose des garanties couvrant les 4 domaines suivants :

- Le décès : versement de capital ou de rentes ;

[14] https ://pro.apicil.com/actualites/loi-evin-prevoyance-definition/

- La dépendance : versement de rente viagère ou d'un capital ;
- L'invalidité : rente complémentaire à la sécurité sociale ;
- L'incapacité : versement d'indemnités journalières.

Pour cela, la Prévoyance fait entrer en jeu le régime de la Sécurité Sociale et les assurances complémentaires qui sont les mutuelles, les assurances et les instituts de Prévoyance qui permettent de compléter les versements des aides sociales

Pour les assurances complémentaires, la Prévoyance collective qui est obligatoire est distinguable de la Prévoyance complémentaire individuelle qui est facultative. Cette dernière permet de compléter les indemnités qui peuvent parfois être insuffisantes selon la situation.

La Prévoyance collective peut intervenir en tant que supplément de versement par rapport au régime obligatoire. Elle concerne un ensemble de salariés et n'est pas tarifiée individuellement mais propose le même tarif pour tous les affiliés du contrat sans discrimination d'âge, du type de contrat etc. C'est un contrat de Branche ou d'entreprise qui est soit obligatoire soit facultatif selon le choix de l'employeur mais qui peut être exigé certaines conventions collectives nationales et par des accords de branches. En tant que cadre, l'employeur est obligé de proposer ce contrat selon la convention nationale des cadres de 1947, tandis que pour un non-cadre ceci est facultatif. En résumé, elle est obligatoire dans trois cas :

- Lorsqu'un employé a plus d'un an d'ancienneté au sein de la société. Dans cette situation, la loi du 19 janvier 1978, aussi appelée loi de « mensualisation » intervient lors d'un arrêt de travail dû à un accident ou une maladie. Les employeurs sont dans l'obligation (sous conditions) de maintenir la rémunération totale ou partielle du salarié. Pour s'assurer de cette obligation, ils peuvent choisir un contrat de Prévoyance collective. C'est pour cette raison que certains assureurs proposent des contrats dits de mensualisation qui couvrent les obligations de l'employeur.
- Pour les cadres, la convention collective nationale de mars 1947 oblige les employeurs à faire profiter de la garantie Décès.
- Lorsque la convention collective l'exige pour les branches ou les entreprises.

Maintenant qu'il a été décrit en quoi consiste la Prévoyance, nous définissons le risque arrêt de travail qui est à l'origine de ce besoin d'assurance.

1.2.2 Définition de l'arrêt de travail

L'arrêt de travail peut être dû à deux types d'évènements : l'incapacité temporaire de travail qui est appelée « incapacité » et l'incapacité permanente de travail qui est nommée « invalidité ». Nous allons nous concentrer sur le risque incapacité dont la définition est donnée : c'est la perte des facultés permettant de poursuivre ses activités en raison d'un accident ou d'une maladie touchant le salarié. L'incapacité est passagère tandis que l'invalidité est due à une dégradation de l'aptitude à travailler à jamais. C'est-à-dire que l'état d'incapacité peut durer 3 ans au maximum, sinon l'individu rentre dans l'invalidité. C'est au médecin de prescrire l'arrêt de travail au salarié ainsi que sa durée lors d'une consultation.

Afin de garantir au salarié une situation financière convenable lors de ces événements, trois acteurs entrent en jeu pour compenser la perte du salaire par un pourcentage de celui-ci : la Sécurité Sociale, l'employeur et les organismes d'assurance vus précédemment.

Notons que tous les salariés ne reçoivent pas la même indemnisation, même pour un motif d'arrêt commun. Celle-ci dépend du statut du salarié, de son organisme de Sécurité Sociale et des assurances souscrites.
Les différents types d'indemnisation de l'arrêt de travail sont expliqués ici.

1.2.3 Indemnisation

1.2.3.1 Indemnisation par la sécurité sociale

La rémunération dépend du motif de l'arrêt : accident du travail, maladie professionnelle, arrêt lié à un accident de la vie privée ou arrêt pour maternité. Elle dépend aussi du secteur d'activité du salarié.
Pour les arrêts liés aux accidents de la vie privée ou à la maternité, l'indemnité journalière est calculée en pourcentage (50%) du salaire journalier de base qui est lui-même calculé à partir de de la moyenne des salaires bruts des trois derniers mois travaillés. Elles sont dues à partir du quatrième jour d'arrêt de travail.
Pour les arrêts liés aux accidents du travail et aux maladies professionnelles, les indemnités journalières sont égales à un pourcentage du salaire journalier qui est lui-même calculé ainsi : le maximum entre 343,07 € et le montant de salaire brut du mois précédent l'arrêt divisé par 30,42. Le pourcentage dépend de la durée de l'arrêt et est limité au salaire journalier diminué de 21%.

1.2.3.2 Indemnisation par l'employeur

Si le salarié est atteint de maladie, d'un accident du travail ou d'une maladie professionnelle, l'employeur a l'obligation de continuer à verser le salaire sous certaines conditions dans un délai de carence de 7 jours. Ce versement diffère en fonction de l'ancienneté du salarié. S'il est ancien d'un an dans l'entreprise, 90% du salaire brut lui est dû pendant les 30 premiers jours et 66,66% les 30 jours suivants. Toutefois, les conventions collectives peuvent venir améliorer ces obligations issues du code du travail.

1.2.3.3 Indemnisation par l'assurance

Le versement des indemnités journalières s'effectue seulement si la durée de l'arrêt dépasse la durée de franchise. L'indemnisation dépend de plusieurs conditions sur le contrat : la garantie, l'ancienneté, la franchise et la durée de versement des prestations.

Voici un schéma résumant l'indemnisation de l'arrêt de travail par ces différents acteurs.

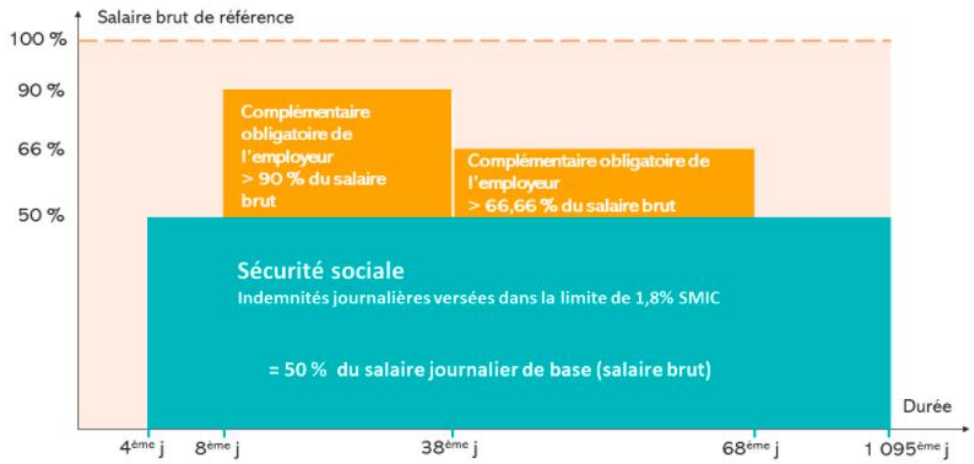


Figure 6 : indemnisation de l'arrêt de travail par les différents acteurs[15]

[15] https://pro.apicil.com/actualites/la-mensualisation-expliquee/

Après avoir décrit l'arrêt de travail et sa gestion par les différents acteurs, nous pouvons introduire la manière dont de ce risque est anticipé par les assurances grâces aux tables règlementaires du BCAC.

Section 1.3 : estimation par les tables réglementaires du BCAC

La loi Evin du 31 Décembre 1989 oblige les organismes d'assurance à provisionner les risques en cours comme le décès, les risques concernant l'intégrité physique, la maternité, le chômage et les risques liés à l'incapacité de travail. Grâce à cette loi qui vient protéger l'assuré, des tables de maintien en incapacité et d'entrée en invalidité provenant du BCAC ont pu être construites en 1993 dans le cadre du provisionnement pour le risque incapacité. Elles furent élaborées sur un grand portefeuille comprenant plusieurs grandes compagnies françaises : AXA, AGF, GAN et UAP. Leur construction s'est faite sur la base d'un modèle paramétrique.

L'arrêté du 28 mars 1996 qui comporte l'article A.331-22[16] du code des Assurances indique qu'il est possible d'utiliser les tables réglementaires du BCAC pour la tarification et le provisionnement. Leur utilisation est encadrée par les articles A. 331-22 du Code des Assurances, A. 931-10-9 du Code de la Sécurité Sociale et A. 212-9 du Code de la Mutualité. Cependant, les compagnies d'assurances peuvent aussi utiliser leurs lois construites par elle mêmes et certifiées par un actuaire reconnu.

Pour comprendre comment fonctionne les lois du BCAC, nous présentons ici la table de maintien en incapacité construite en 1996 et prolongée en 2010 compte tenu de la réforme des retraites.

La table d'entrée en incapacité ci-dessous considère 10 000 personnes entrées en incapacité à l'âge de 20 ans et indique pour chaque âge et selon le temps passé en incapacité en mois, combien de personnes sont encore en incapacité.

Un extrait de cette table est présenté. Elle est coupée à l'âge de 40 ans et à la durée de 19 mois.

	MOIS																			
ÂGE	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
20	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
21	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
22	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
23	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
24	10 000	2 931	1 848	1 215	894	657	478	343	291	256	217	183	166	143	130	121	114	105	95	91
25	10 000	3 080	2 001	1 345	997	739	536	382	327	289	251	216	195	172	159	149	140	129	116	113
26	10 000	3 177	2 112	1 461	1 087	812	591	431	372	325	285	249	226	201	186	171	161	150	137	129
27	10 000	3 251	2 180	1 540	1 156	869	643	476	407	360	320	285	263	237	222	207	192	179	168	159
28	10 000	3 298	2 243	1 600	1 209	915	688	524	448	400	359	322	297	270	255	238	222	210	199	189
29	10 000	3 348	2 273	1 640	1 246	956	726	559	476	425	384	352	327	298	280	262	247	233	220	208
30	10 000	3 386	2 275	1 659	1 264	964	744	583	494	439	396	363	338	308	287	267	252	240	227	214
31	10 000	3 388	2 228	1 618	1 249	965	756	595	501	449	406	375	347	318	295	276	261	250	236	223
32	10 000	3 433	2 238	1 617	1 254	975	772	612	522	468	421	388	357	325	302	279	264	252	235	222
33	10 000	3 466	2 235	1 627	1 260	983	782	628	540	484	431	395	364	332	310	286	270	256	238	223
34	10 000	3 567	2 298	1 684	1 321	1 033	828	684	597	535	477	436	401	366	344	319	298	282	265	247
35	10 000	3 645	2 331	1 705	1 357	1 082	876	732	647	586	528	481	443	402	377	351	331	309	294	275
36	10 000	3 701	2 390	1 747	1 390	1 106	905	771	682	617	560	508	469	428	397	370	347	323	308	287
37	10 000	3 822	2 458	1 804	1 430	1 148	932	801	704	635	579	526	487	443	406	379	357	335	319	298
38	10 000	3 958	2 526	1 851	1 479	1 193	980	841	739	671	616	564	521	477	439	411	384	358	340	319
39	10 000	4 035	2 600	1 923	1 541	1 266	1 055	915	807	739	680	623	572	530	486	455	427	400	381	364
40	10 000	4 073	2 652	1 973	1 575	1 303	1 097	965	853	783	719	659	607	565	521	490	458	428	404	384

Figure 7 : loi de maintien en incapacité[17]

[16] https ://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006788054/2005-12-16
[17] http://www.ressources-actuarielles.net/bcac

Cette table qui a été créée en 1996 est calibrée sur des survenances qui datent et qui ne reflètent pas forcément les sinistres actuels. De plus, elle est construite d'après des portefeuilles de société d'assurance qui n'avaient pas l'habitude de proposer des contrats avec des franchises courtes. Elle ne modélise donc pas bien l'absentéisme actuel et n'est pas pertinente pour estimer la durée des arrêts de court terme.

Les tables du BCAC sont très génériques et axés uniquement sur les facteurs âge et ancienneté. Il serait intéressant de tester de nouveaux facteurs qui expliquent la dérive de l'absentéisme par une approche individuelle.

Nos propres lois sont donc élaborées afin de tarifier les arrêts de travail liés à l'incapacité. Celles-ci peuvent être optimisées grâce aux nombreuses données historiques de qualité qui proviennent de la Déclaration Sociale Nominative qui est introduite maintenant.

Section 1.4 : la DSN

1.4.1 Définition et évolution

Depuis la loi du 22 mars 2012, a été instaurée la DSN. Elle est mise en œuvre depuis 2017 de manière progressive et remplace la DADS-U qui est la déclaration annuelle des données unifiées. La DADS-U obligeait les employeurs à déclarer chaque année les informations suivantes :

- L'identité de l'employeur ;
- L'emploi du salarié et le contrat de travail ;
- La période de travail ;
- Les salaires reçus ;
- Les cotisations d'assurance vieillesse et d'assurance maladie.

Cependant, certains employeurs ne peuvent pas déclarer via le système DSN et passe encore par la DADS-U :

- Des particuliers employeurs ;
- Des employeurs de la fonction publique ;
- Des entreprises localisées dans des zones géographiques non visées par le système de la DSN (Monaco, certaines collectivités d'Outre-Mer, etc…) ;
- Des employeurs dont les salariés sont hors périmètre de la DSN ;
- Des établissements qui viennent d'intégrer la DSN et qui n'ont donc pas encore pu communiquer les informations des organismes complémentaires en DSN.

Contrairement à la DADS-U, la DSN est un flux de données provenant de déclarations mensuelles. Elle est à charge des employeurs et se fait à partir du logiciel de paie en ligne qui contient toutes les informations du salarié, bien plus nombreuses que celles contenues dans le dernier système de déclaration.

Ses objectifs sont multiples. La DSN permet de centraliser les informations des salariés et permet donc de simplifier la gestion des salariés. Grâce à cela, nous observons une facilitation des calculs et des paiements des

cotisations sociales, mais aussi une facilitation de la transmission des informations détaillées des salariés et des évènements au travail à tous les organismes qui en ont besoin comme Pôle Emploi, l'assurance maladie, l'Urssaf, Argic -Arrco, et les organismes complémentaires de santé.

Les déclarations suivent des règles précises. Il est à noter que pour chaque établissement du secteur privé, du régime général et du régime agricole, et donc pour chaque numéro de SIRET, une déclaration doit être déposée.
Il y a différents types de déclarations et des délais sont instaurés pour toutes. En plus des déclarations classiques, les évènements inhabituels sont à communiquer dans des déclarations évènementielles. Selon la taille de l'entreprise (moins de 50 salariés ou plus de 50 salariés), la déclaration doit être faite le 15 du mois ou le 5 du mois. Lors d'un arrêt de travail ou d'une fin de contrat par exemple, la déclaration doit être effectuée dans les cinq jours suivant l'évènement.

Toutes ces règles proviennent d'une norme : la norme NEODeS qui est la Norme d'Echange Optimisée des Données Sociales. Elle est à l'origine de la DSN et est décrite dans le cahier technique de 2020[18]. Dans celui-ci, sont précisées pour chaque donnée, des rubriques avec des modalités et des contrôles appliqués à ces données. En fait, c'est une « grammaire » qui est applicable par tous les déclarants. Elle définit une structure de la DSN très précise. Nous y trouvons toutes les règles relatives aux déclarations, leurs constitutions ou des déclarations ou des signalements d'évènements comme les ruptures de contrats, les arrêts de travail etc. Elle décrit de manière technique les informations de la fiche de paie d'un salarié à transmettre : le salarié, le contrat, la rémunération, les montants de cotisations etc.
Cette norme est la conséquence de simplification du flux de la DSN et contient seulement les informations des fiches de paie et des données RH. Elle simplifie la centralisation des données administratives par son automatisation.

L'évolution de la DSN jusqu'à sa généralisation est présentée sur le schéma ci-dessous.

[18] https ://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2020.pdf

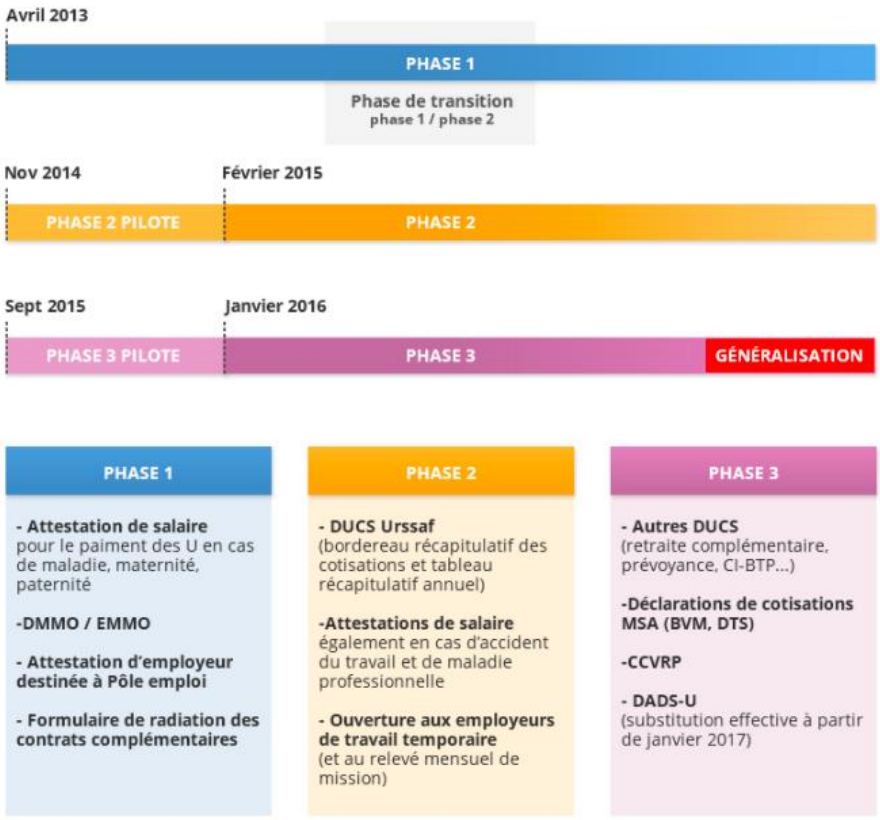


Figure 8 : calendrier de la DSN[19]

1.4.2 Apport de la DSN pour l'estimation du coût de l'arrêt de travail

Avant la DSN, les assureurs connaissaient peu les salariés d'une entreprise souscriptrice d'un contrat de Prévoyance collective. En Prévoyance, nous sommes sur un risque de fréquence faible contrairement à la Santé où l'assureur prenait connaissance du salarié seulement en cas de réalisation d'un risque. Ainsi, l'assureur n'effectuait pas d'affiliation dans son système d'informations des salariés pour un contrat de Prévoyance collective. Les systèmes d'informations des assureurs n'avaient pas non plus été créés pour traiter des DADS-U. Le changement provoqué par la DSN est que l'organisme reçoit directement les flux DSN et ne peut pas demander à l'émetteur une information déjà transmise. Depuis la DSN, les entreprises transmettent directement à l'assureur les éléments pour le calcul et le paiement des cotisations dues au titre de la Prévoyance.

De plus, la DSN est particulièrement enrichissante au niveau des informations que nous pouvons recueillir. Cela constitue une révolution dans le monde des assurances. Nous l'avons compris, elle a une volumétrie particulièrement importante qui nous apporte un détail très fin à la maille du salarié. Les informations du salarié, de ses contrats, de ses arrêts et d'autres éléments encore sont très précis. Les comportements déclaratifs peuvent également nous apprendre beaucoup sur les individus, les déclarations étant variées.

Aussi, la DSN nous offre une vision récente des informations des salariés, et donc de meilleure fiabilité car elle est mise à jour mensuellement.

[19] Source : cegedim-srh.com

Enfin, dans le cadre d'une étude sur l'absentéisme, nous avons à présent bien plus de vision sur les salariés, tant sinistrés que non sinistrés, et sur tout type d'arrêts quelques soient leurs durées, tant les arrêts longs que les arrêts courts. Alors qu'auparavant, les bases des systèmes de gestion opérants de la Prévoyance collective nous offraient une vision plus limitée, qui comprenait seulement les salariés indemnisés et donc sinistrés, et uniquement les arrêts qui dépassent la franchise.
Dans ces bases d'indemnisation, il peut aussi y avoir un phénomène de sous-déclaration des arrêts par l'entreprise dès lors qu'il y a une subrogation. Cela arrive lorsque l'entreprise n'attend pas d'être remboursée par l'assureur pour verser les indemnités dues par le contrat de Prévoyance et que la durée de l'arrêt est très légèrement supérieure à la durée de la franchise.

De ces faits, la DSN permet ainsi de mieux répondre aux besoins pour une étude sur l'absentéisme.

A présent, nous pouvons présenter notre étude qui utilise ces données DSN retravaillées qui permettent d'établir une étude assez complète sur l'absentéisme.

Partie 2 : les données

Dans cette partie, l'étude sera d'abord présentée par son périmètre et les données disponibles. Ensuite, les bases de données seront retraitées et nettoyées. Il sera alors possible de présenter des statistiques descriptives et des études de corrélations pour mieux comprendre nos bases.

Section 2.1 : présentation de l'étude

2.1.1 Périmètre de l'étude et données cibles

L'étude porte sur les branches « aide à domicile » et « propreté » car elles présentent un enjeu important pour le portefeuille. AG2R couvre les arrêts de très court terme sur ces deux CCN, il est donc indispensable de maitriser ce risque notamment en tarification. Ce sont deux branches historiques du portefeuille qui constituent un chiffre d'affaire considérable sur les arrêts de travail et un volume d'arrêts conséquent. Elles sont aussi particulières dans leur population sous risque qui est constituée de femmes dans sa grande majorité, de tailles d'entreprises très diverses et de la présence importante de salariés en temps partiel. Ces branches présentent donc une typologie de risque très spécifique.

Ensuite, rappelons que notre objectif est de modéliser une loi de fréquence qui permet de prédire la fréquence d'arrêts d'un salarié pour un contrat donné au cours d'une année, puis, une loi de maintien qui modélise sa durée en jours.

L'ensemble des informations détaillées qui proviennent de la DSN est mis à profit afin de construire ces lois au niveau individuel.

Pour les modélisations de ce risque arrêt de travail, nous nous basons donc sur les données de la DSN des années 2018 et 2019. Les années précédant 2018 ne sont pas prises en compte car les données ne sont pas de bonne qualité sur celles - ci du fait de la progressivité dans l'adaptation au système DSN par les entreprises. Les années 2020 et 2021 ne sont pas retenues non plus en raison de la crise sanitaire Covid 19 qui a eu de grosses conséquences sur l'absentéisme. En effet, si la modélisation de l'arrêt de travail est effectuée sur cette période, les résultats seraient fortement biaisés et ne représenteraient pas la réalité du risque dans les temps normaux.
Ce choix de périmètre restreint n'est pas sans conséquence car nous ne pouvons capter que les durées d'arrêts de deux ans maximum. En plus de cette restriction, afin d'éviter les problèmes de censures et de troncatures sur les arrêts et leur durée, nous modélisons seulement les arrêts ayant leur survenance sur la période d'observation fixée du 01/01/2018 au 31/12/2019 et ayant une durée inférieure à 6 mois. Seulement les arrêts de courte durée sont étudiés. Néanmoins, pour modéliser les arrêts de longue durée, un raccordement de loi

peut être effectué avec des lois construites sur les données issues des systèmes de gestion opérants (outil de gestion de versement des prestations arrêt de travail et outil de paramétrage des contrats) et non de la DSN, la faiblesse de celles-ci portant plutôt sur les arrêts courts. Ce dernier point ne sera pas présenté dans le mémoire.

Mais encore, les arrêts ne sont pas tous retenus pour notre étude. Nous nous focalisons uniquement sur les arrêts liés aux motifs suivants :
- maladie ;
- congé suite à un accident de trajet ;
- congé suite à une maladie professionnelle ;
- congé suite à un accident de travail ou de service.

Les arrêts liés à la maternité, la paternité ou l'accueil de l'enfant, le temps partiel thérapeutique et l'adoption sont exclus de l'étude. Ces arrêts sont trop spécifiques et sont plus longs que pour les autres motifs. Au risque de biaiser nos modélisations, ils sont éliminés de la base.

Enfin, nous nous restreignons au périmètre de la population active. Ce sont seulement les salariés de plus de 18 ans et de moins de 67 ans qui sont sélectionnés.

Nous présentons maintenant les données disponibles que nous avons à disposition dans une base qui est à retraiter.

2.1.2 Présentation des données disponibles

Nous avons à notre disposition une base, nommée DVA, contenant les informations du salarié, de ses arrêts, de ses contrats et des entreprises déclarantes. Un processus a été mis en place pour parvenir à la construction de cette base qui recueille les données de la DSN, parfois retravaillées pour qu'elles soient exploitables dans le cadre d'une étude sur l'absentéisme. Ainsi, une méthodologie particulière a été suivie pour obtenir cette base de données à partir de 29 tables de la DSN. Voici l'architecture de la construction de la base :

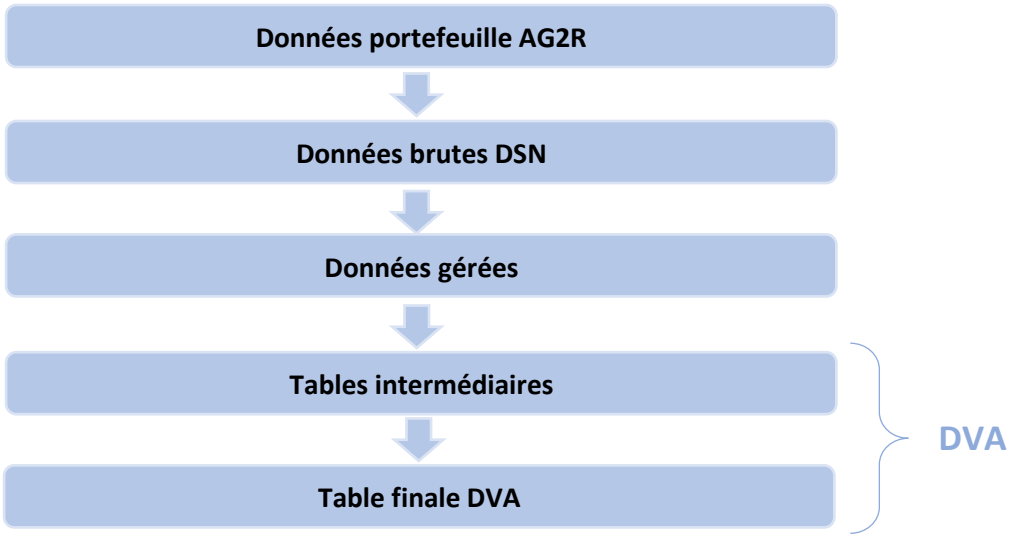


Figure 9 : schéma d'alimentation du DVA

Explication de ce Schéma

Le premier bloc contient les données du portefeuille AG2R. Puis, en fonction de la demande, il alimente la base des données brutes de la DSN où nous retrouvons les données du salarié directement, les fiches de paie etc. Elles-mêmes alimentent de manière hebdomadaire les données gérées qui sont les données brutes de la DSN mais avec les nomenclatures des champs changées afin que les tables soient exploitables. Ces données gérées donnent naissance au DVA (Données à Valeur Ajoutée).
La création du DVA se fait via des règles de gestion sur les variables pour chaque table pour ensuite les agréger et en obtenir une seule exploitable : la table finale DVA historisée pour les années de déclaration allant de 2018 à 2021. Celle-ci est donc produite grâce à plusieurs tables d'alimentations : la table des salariés, la table des arrêts et la table des rémunérations.
Notons que ces bases sont à la maille SIREN, salarié, année de déclaration DSN, arrêt. Pour chaque salarié, nous retrouvons donc dans la base une première ligne de déclaration classique (non événementielle) du dernier mois de l'année avec les informations caractérisant le salarié et son contrat. Les lignes suivantes représentent les déclarations d'arrêts avec une ligne par arrêt.

Maintenant que nous avons accès à cette base, nos besoins pour l'étude sur les arrêts de travail sont ciblés. Si des variables que nous avons besoin sont absentes, elles sont rapatriées par la suite avec les tables de la DSN directement. Nous énumérons les variables provenant de la DSN qui nous intéressent et que nous retrouvons dans le DVA pour la plupart en **annexe 3**. Elles sont détaillées dans le cahier technique de la DSN[20] à l'aide du code associé aux modules inscrits dans la première colonne (tables) du tableau en **annexe 3**. Ces variables détiennent les informations que nous voulons obtenir sur l'entreprise et l'établissement de travail, l'individu, la pénibilité de l'emploi de l'assuré, les détails du contrat, la rémunération, l'arrêt de travail, le lieu du travail, les motifs de suspension de contrat, les cotisations individuelles et celles de l'établissement.

Dans le paragraphe suivant est détaillée la plus-value de la DSN par rapport aux données issues des systèmes de gestion opérants pour notre étude sur l'absentéisme. Cela permet de comprendre la réelle motivation de construire une base de données qui contient les informations de la DSN.

2.1.3 La plus-value de la DSN par rapport aux données issues des systèmes de gestion opérants

L'exploitation des systèmes de gestion opérants pour notre étude sur les arrêts courts ne satisfait pas nos exigences. Ce sont donc les données de la DSN de plus en plus fiables qui sont utilisées pour les travaux de tarification. Les avantages de l'utilisation de ces données (énoncés dans la section 1.4.2) par rapport à celles des systèmes de gestion opérants sont démontrés ici.

Ce nouveau flux de données nous permet d'avoir des informations sur les salariés sinistrés et non sinistrés, tandis que les systèmes de gestion (infocentre) se limitent aux salariés sinistrés (car indemnisés par l'assurance). Pour mettre en évidence ce manque d'informations de l'infocentre, les lois qui ont été créées grâce aux données de l'infocentre sont comparées avec les lois du BCAC.

[20] https ://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2020.pdf

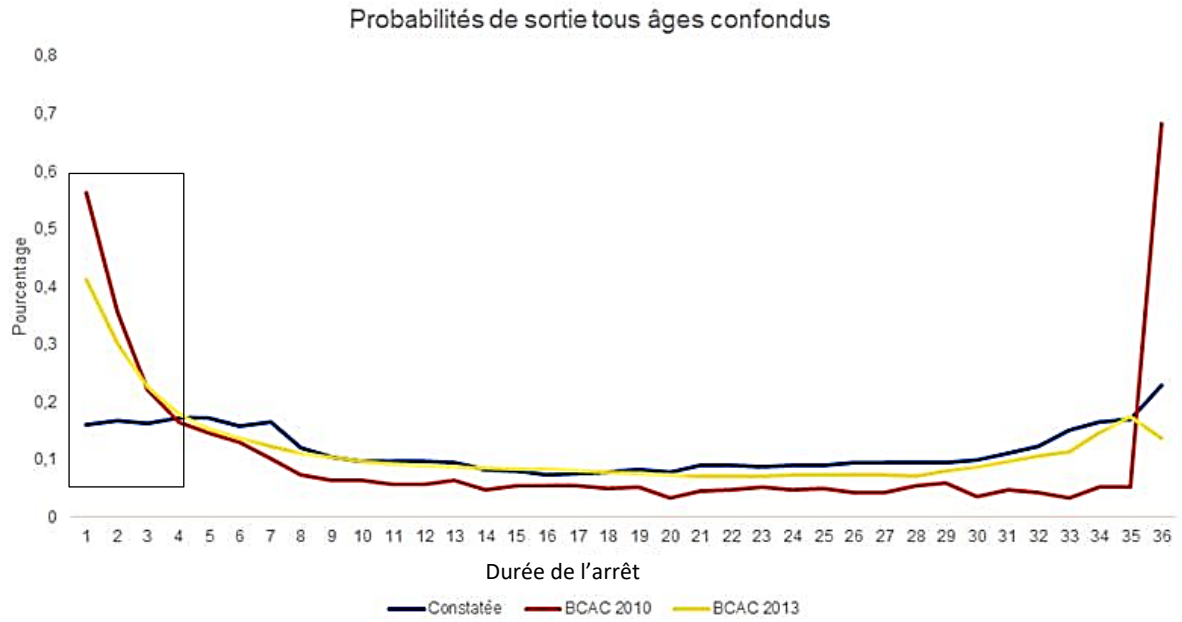


Figure 10 : comparaison de la loi construite sur les données de l'infocentre avec la loi du BCAC

Ce graphique indique les probabilités de sortie de l'état d'incapacité en fonction de l'ancienneté dans l'arrêt. Il vient accentuer les failles de l'infocentre. Comme nous pouvons le constater sur ce graphique, durant les quatre premiers mois d'ancienneté, d'après le BCAC il y a beaucoup d'individus qui sortent alors que d'après la loi créée grâce aux données de l'infocentre (en bleu), nous observons qu'il y en a moins : un taux de 0,15 pour la loi constatée contre 0,4 pour le BCAC 2013 pour une ancienneté d'un mois. Cet écart entre les deux courbes met donc en évidence le manque d'informations sur les arrêts courts, pendant la période de franchise. Toutefois, comme le prouve le croisement des courbes à 4 mois d'ancienneté et leur rapprochement après 4 mois dans l'arrêt, les données de l'infocentre suffisent à la modélisation des arrêts longs.

Un autre avantage de la DSN, pas des moindres, est la présence d'information plus fine et plus riche sur les salariés que les bases de l'infocentre. Pour illustrer le fait que la DSN est plus riche en données que l'infocentre, nous listons les variables qui sont présentes dans la DSN et qui sont manquantes dans les données de l'infocentre. La comparaison se fait avec la table des encours des arrêts qui contient les personnes qui perçoivent actuellement une rente et sont indemnisées (cette table est utilisée pour les comptes de résultat et provient de l'infocentre).

Voici donc une liste non exhaustive d'informations qui manqueraient pour notre étude sur l'absentéisme si on se base sur l'infocentre :

Entreprise
- Effectif moyen
- Code du lieu de travail de l'assuré

Penibilité
- Facteur d'exposition

Salarié et contrat de travail
- Date de début
- Statut catégoriel retraite complémentaire
- Date de fin prévisionelle
- Temps partiel/temps plein
- Type de régime de base de l'assuré
- Type de régime spécial
- Identifiant du régime de base vieillesse
- Motif de l'embauche
- Niveau de formation de l'individu

Rémunération
- Nombre d'heures concernées

Arrêt de travail
- Date de fin prévisionnelle de l'arrêt
- Motif de la reprise (temps partiel, …)

Fin du contrat
- Date de fin de contrat
- Motif de la rupture du contrat

Cependant, certaines données des tables de l'infocentre restent indispensables pour certaines études car elles contiennent par exemple les différentes informations concernant les indemnisations des salariés, ce que nous n'avons pas dans les DSN.

Pour pouvoir utiliser les données qui sont à notre disposition, il est nécessaire de les retraiter.

Section 2.2 : retraitements

Plusieurs retraitements de la base sont nécessaires afin d'obtenir une base de données fiable et exploitable pour notre étude. L'ensemble des retraitements réalisés est présenté.

2.2.1 Les retraitements préliminaires

Dans la base, la variable « idsalarie » n'identifie pas de manière unique un salarié mais plutôt les contrats du salarié car cet identifiant change pour chaque contrat. Nous cherchons donc à obtenir un identifiant du salarié qui le rend unique et qui pourra nous permettre d'avoir des résultats non biaisés pour l'étude. Pour dédoublonner les salariés, nous utilisons donc une table de correspondance qui contient l'« idsalarie», le numéro de Sécurité Sociale du salarié, son nom, son prénom et sa date de naissance. Ainsi, le salarié peut être défini par son NIR de manière unique.

De nouvelles variables jugées intéressantes pour notre étude sont créées :

- Des tranches de rémunération annuelle sont construites pour avoir une information moins fine sur les rémunérations et pour pouvoir analyser par la suite l'influence de tranches de rémunération sur l'arrêt.
- L'ancienneté du salarié dans l'entreprise en mois et en années est calculée en prenant la différence entre la date de déclaration du sinistre et la date de début de contrat, car dans la base DVA, celle-ci n'est pas calculée pour les lignes sans arrêts.
- Le ratio d'exposition est calculé : somme de la durée de présence en entreprise (en jours) /365 (il est vérifié que le ratio est toujours compris entre 0 et 1).
- La variable du libellé de la nature du contrat est recatégorisée car les modalités sont trop nombreuses et peu exploitables.

Modalités de la variable	Catégories
Contrat de travail à durée déterminée de droit public	CDD
Contrat de mission (contrat de travail temporaire)	CDD
Convention de stage (hors formation professionnelle)	CDD
Contrat d'engagement éducatif	CDD
Autre nature de contrat, convention, mandat	CDD
Mandat social	CDD
Contrat de travail à durée indéterminée de droit privé	CDI
Contrat à durée indéterminée intérimaire	CDI
Contrat à durée indéterminée intermittent	CDI
Contrat de travail à durée indéterminée de chantier ou d'opération	CDI
Contrat de travail à durée indéterminée de droit public	CDI

- La variable indiquant le code délégataire de l'organisme d'adhésion à la Prévoyance contient 33 modalités et des valeurs manquantes. Cette variable n'est donc pas exploitable. Nous créons donc une nouvelle variable qui indique « gestion déléguée » lorsque la variable contient un code et « gestion directe » sinon.
- La variable contenant le libellé du statut conventionnel du salarié est recatégorisée et les modalités qui sont peu présentes dans la base sont regroupées dans la catégorie « Autres ».

Modalités de la variable	Catégories
Employé administratif d'entreprise, de commerce, agent de service	Employé administratif d'entreprise, de commerce, agent de service
Profession intermédiaire (technicien, contremaitre, agent de maitrise, clergé)	Profession intermédiaire (technicien, contremaitre, agent de maitrise, clergé)
Ouvriers qualifiés et non qualifiés y compris ouvriers agricoles	Ouvriers qualifiés et non qualifiés y compris ouvriers agricoles
Agent de la fonction publique territoriale	Agent de la fonction publique territoriale
Autres cadres au sens de la convention collective (ou du statut pour les régimes spéciaux)	Cadre
Cadre dirigeant (votant au collège employeur des élections prud'homales)	Cadre
Agriculteur salarié de son exploitation	Autres
Agent de la fonction publique d'Etat	Autres

- La variable contenant le libellé du statut du salarié est recatégorisée.

Modalités de la variable	Catégories
Non cadre	Non cadre
Pas de retraite complémentaire	Non cadre
Retraite complémentaire ne définissant pas de statut cadre ou non cadre	Retraite complémentaire ne définissant pas de statut cadre ou non cadre
Cadre (article 4 et 4bis)	Cadre
Extension cadre pour retraite complémentaire	Cadre

Après avoir effectué ces retraitements sur les variables déjà présentes dans la base, nous rapatrions d'autres variables qui nous intéressent.

2.2.2 Enrichissement par des variables calculées et des variables externes

La DSN contient des informations au niveau assuré × contrat × arrêt.

Au niveau de l'assuré, les coordonnées géographiques nous permettent de calculer la distance moyenne par entreprise entre le lieu de travail et le lieu de résidence de l'assuré à l'aide des longitudes et latitudes

renseignées sur l'entreprise et le salarié. Celle-ci peut influencer les arrêts de travail d'un salarié et est donc intéressante pour notre analyse.

Au niveau de l'entreprise, le code APE nous permet de classer les entreprises dans des secteurs d'activité lourds, moyens et légers, ce qui peut avoir un impact significatif sur les arrêts de travail.
Grâce aux tables de l'INSEE[21], il est aussi possible d'obtenir les informations sur le degré de densité de la commune (allant de 1 à 4 pour la moins peuplée à la plus dense) et sur les revenus médians par commune. Ces critères pourraient avoir une influence sur les arrêts de travail.

Au niveau du contrat, les données techniques comme la franchise n'ont pas pu être récupérées. C'est une première limite à notre étude.

Maintenant que les variables qui nous intéressent sont présentes dans notre base, il faut faire attention aux valeurs manquantes et aberrantes qu'elles peuvent contenir.

2.2.3 Valeurs manquantes et aberrantes

Les traitements des variables qui présentent des valeurs manquantes ou/et aberrantes sont détaillés ici.

- Les valeurs des rémunérations négatives sont rendues positives. Cette décision est prise en regardant la cohérence des emplois avec les montants de rémunération. Pour des salariés cadres en CDI, il est illogique d'avoir une rémunération négative.

- La variable indiquant la date de fin prévisionnelle (prescrite par le médecin) de l'arrêt de travail est toujours remplie tandis que la date de dernier jour travaillé est souvent manquante. Or, il est préférable d'avoir la date de dernier jour travaillé réelle plutôt que la prévisionnelle. Nous créons donc une nouvelle variable prenant la valeur du dernier jour travaillé si cette variable est remplie, prenant la date prévisionnelle sinon. Elle est nommée « dtfinAT ».

- La variable « nbjourpresenceentreprise » indique la durée de présence d'un salarié en entreprise pour un certain contrat sur un an. Pourtant, des valeurs négatives et supérieures à 365 sont observées. Elle est recalculée de cette manière : nous considérons pour chaque année DSN la date de début égale au maximum entre la date de début de contrat et le 01/01/année DSN. Puis, si la date de fin de contrat n'est pas renseignée alors nous retenons comme date de fin le 31/12/année DSN, sinon nous retenons le minimum entre la date de fin du contrat renseignée et le 31/12/année DSN. Finalement, le nombre de jours de présence en entreprise est égale à : date de fin - date de début + 1. Nous ajoutons « +1 » car la date de fin du contrat est incluse dans la présence en entreprise d'après le cahier technique de la DSN.
  (Avant cela, nous vérifions que la date de début de contrat est toujours inférieure à la date de fin de contrat et que celle-ci est toujours supérieure au 01/01 de l'année DSN correspondante.)

- Nous supprimons les contrats dont les arrêts ne sont pas compris dans la période de présence en entreprise et ceux qui génèrent plus d'arrêts que de jours de présence en entreprise.

[21] Niveau de densité par commune : https ://www.insee.fr/fr/information/2114627
Revenu médian par commune : https ://www.insee.fr/fr/statistiques/1293130

- Le statut étranger du salarié, le motif d'embauche et les motifs de fin de contrat ont une majorité de valeurs manquantes. Ces variables ne sont pas utilisées dans les modélisations à venir.

Afin de construire des bases d'apprentissage pour les modélisations, d'autres retraitements sont effectués.

2.2.4 Bases d'apprentissage

Les champs qui sont obligatoirement renseignés dans les DSN sont les données identifiantes qui permettent de faire le lien entre plusieurs déclarations. Les quatre blocs indispensables à renseigner sont :

1. Le bloc entreprise qui est identifié par le SIREN ;
2. Le bloc établissement qui est caractérisé par le NIC ;
3. Le bloc individu qui est défini par le NIR, le nom de famille, le prénom et la date de naissance de l'assuré ;
4. Le bloc contrat qui est identifié par le numéro de contrat et la date de début de contrat.

Dans la DSN, la maille la plus fine sur laquelle nous pouvons donc nous baser pour avoir des informations sur les salariés et ses arrêts de travail est la maille contrat.

Les mailles choisies pour la base d'apprentissage sont différentes selon la loi que nous voulons modéliser. Deux bases d'apprentissage différentes sont construites. Pour la loi de fréquence, une base à la maille salarié $\times$ contrat est élaborée afin de prévoir la fréquence d'arrêts selon les caractéristiques du salarié affilié à un contrat particulier. Pour la loi de maintien, une base à la maille arrêt est exploitée, car en fonction de chaque type d'arrêt avec ses différents critères, la durée de l'arrêt est prédite.

Une première base à la maille arrêt est d'abord construite, puis, une base d'apprentissage pour la loi de fréquence est conçue à partir de cette dernière.

2.2.4.1 Création de la base à la maille arrêt

Tout d'abord, les arrêts sont dédoublonnés afin d'obtenir un arrêt par ligne en utilisant les corrections suivantes si nécessaire :

- Lorsqu'il y a une même date de dernier jour travaillé « dtdernierjour » pour deux dates de fin d'arrêt de travail « dtfinAT » différentes, nous retenons la date de fin d'arrêt de travail maximum par « anneedsn », « idsalarie » et « dtdernierjour » ;
- Un contrat d'un salarié possède parfois la même date de fin d'arrêt de travail pour deux dates de dernier jour différentes. Dans ce cas-là, c'est le minimum de « dtdernierjour » par « dtfinAT » qui est pris ;
- Lorsque la période d'un arrêt est comprise dans une autre : seul l'arrêt dont la période englobe l'autre est retenu. L'autre arrêt est considéré comme un doublon et est supprimé ;
- Les arrêts qui sont doublonnés grâce à nos retraitements précédents sont supprimés.

Une vérification est effectuée afin de s'assurer que la date de fin d'arrêt de travail est toujours supérieure à la date de début. C'est bien le cas.

Un recalcul des variables liées à l'arrêt est effectué suite au dédoublonnage des arrêts précédemment décrits :

- Pour chaque arrêt, la durée de l'arrêt est recalculée en jours en faisant la différence entre la date de reprise du travail et la date de dernier jour travaillé ;
- Le délai de rechute est recalculé pour chaque arrêt en faisant la différence entre la date de reprise du travail d'un arrêt et la date de début d'arrêt de la rechute suivante ;
- Une variable « rechute » est créée. Elle indique 1 si le délai de rechute est inférieur ou égal à 7, sinon 0.

A présent, la base est filtrée sur les arrêts qui ont une durée inférieure à 6 mois.

2.2.4.2 Création de la base à la maille salarié × contrat

Afin de créer la base à la maille salarié × contrat qui servira à construire notre loi de fréquence, certains traitements sont à effectuer.

- Toutes les variables qui caractérisent l'arrêt (motif de l'arrêt de travail, date de l'arrêt etc) sont retirées car nous voulons agréger les informations au niveau du salarié et de son contrat ;
- Par contrat, nous calculons le nombre d'arrêts (après le traitement de rechutes expliqué dans 2.2.4.3), la durée moyenne des arrêts, la durée totale des arrêts et le nombre de rechutes ;
- Le montant de rémunération mensuel et le salaire journalier sont calculés pour chaque arrêt et ils sont parfois inégaux aux différentes dates de déclaration des arrêts pour un même contrat. Par exemple, le montant de rémunération mensuel n'a pas forcément les mêmes valeurs pour des dates de déclaration distinctes. Nous calculons donc le maximum, le minimum et la moyenne de ces montants par « anneedsn » et « idsalarie », afin d'avoir l'information par contrat d'un salarié ;
- Une variable cible est construite. Elle indique la fréquence d'arrêts par contrat : le nombre d'arrêts divisé par l'exposition du salarié dans son contrat.

Des règles de gestion sont aussi fixées pour éradiquer les rechutes des bases d'apprentissage.

2.2.4.3 Règles de gestion

Dans la base DVA, chaque potentielle rechute est considérée comme un arrêt et constitue une ligne supplémentaire dans la base. De plus, le délai de rechute entre deux arrêts n'est pas calculé par arrêt mais par contrat en prenant le délai écoulé entre le premier et le dernier arrêt d'un contrat d'un salarié. Des règles de gestion sur les rechutes sont donc mises en place afin de corriger ces deux constats.

Pour cela, le délai de rechute est recalculé pour chaque arrêt. C'est-à-dire, pour chaque ligne de la base, nous calculons la différence entre la date de reprise du travail après un arrêt et la date de début d'arrêt de celui qui suit. Lorsque le délai entre deux arrêts successifs est inférieur ou égal à 7 jours, l'arrêt qui suit est considéré comme une rechute, sinon il est considéré comme un autre arrêt.
Maintenant que les arrêts qui sont des rechutes sont identifiés, toutes les rechutes sont agrégées ensemble pour ne former qu'un seul arrêt. Ainsi, lorsque les arrêts sont identifiés comme des rechutes et ont le même motif d'arrêt de travail, cela est résumé en une seule ligne l'arrêt. Nous remontons les informations de la dernière rechute comme la date de fin d'arrêt de travail et le motif de reprise. Aussi, les durées d'arrêts du premier arrêt et des rechutes qui suivent sont sommées. Une base avec seulement des arrêts sans rechutes est ainsi obtenue.

Les bases étant désormais retraitées, les variables d'apprentissage sont sélectionnées pour les modélisations.

2.2.4.4 Variables pour la modélisation

Les variables retenues ne sont pas les mêmes pour modéliser la durée et pour modéliser la fréquence. Pour pouvoir utiliser les variables catégorielles dans l'implémentation des modèles, elles sont transformées en variables binaires. La table finale ainsi créée contient donc les variables quantitatives sélectionnées et les variables catégorielles sélectionnées binarisées.

Les variables qui répètent l'information contenue dans d'autres variables de manière totale ou quasi-totale (cf étude de corrélations : section 2.4) sont éliminées et les variables catégorielles ayant trop de modalités sont éliminées aussi car elles sont non exploitables.

Variables retenues pour le modèle de fréquence

Variables numériques	Modalités prises
agesalariedsn : âge du salarié	Entiers dans [18 ; 67]
nbsalglobalannee : nombre de salariés par entreprise par année	Entiers dans [1 ; 14 675]
pourcentsalglobalanneefemme : pourcentage de femmes dans l'entreprise	Pourcentage entre 0 et 99,82%
pourcentsalglobalanneehomme : pourcentage d'hommes dans l'entreprise	Pourcentage entre 0 et 100%
mtremunerationsalarieannuelle : montant de rémunération annuel brut du salarié pour un contrat	Réels dans [0 ; 1 423 446]
mttremuneration_mean : montant de rémunération mensuel brut moyen par contrat d'un assuré	Réels dans [0 ; 1 423 446]
distance_dom_trav : distance du lieu de travail au domicile du salarié	Réels dans [0 ; 9 708,24]
nbmoisanciennete : ancienneté du salarié dans l'entreprise en mois	Réels dans [0 ; 768,23]

Variables catégorielles	Modalités prises
trancheetablissement : taille de l'établissement	[1 à 10 salariés ; 11 à 50 salariés ; … ; plus de 300 salariés]
trancheage : tranches d'âge	[16 à 19 ans ; 20 à 24 ans ; … ; 60 à 67 ans]
nomregion : nom de la région du lieu de travail	[ILE-DE-France ; RHONE-ALPES ; PAYS DE LA LOIRE; …]
lbsexe : sexe	[Homme ; Femme]
libmotiffinctrt : motif de fin de contrat	[rupture conventionnelle ; licenciement pour faute grave ; ...]
tranchemtremunerationsalarieannuelle : tranches de montant de rémunération annuel brut du salarié pour un contrat	[salaire < 1000 ; salaire 1000-2000 ; … ; salaire >= 40000]
libnaturecontrat : libellé de la nature du contrat	[Contrat de mission (contrat de travail temporaire) ; Contrat à durée indéterminée intérimaire ; …]
cddelegataireprevadh : gestion déléguée ou directe	[gestion directe ; gestion déléguée]
libcdstatutrcsalarie : libellé du statut du salarié	[non-cadre ; cadre (article 4 et 4bis) ; …]
libstatutconventionnelsalarie : libellé du statut conventionnel du salarié	[artisan ou commerçant salarié de son entreprise ; agent de la fonction publique d' Etat]
modaliteexercice : temps plein ou temps partiel	Entiers dans [20 ; 10 ; 99 ; 41] ([Temps plein ; Temps partiel ; non applicable ; Temps partiel de droit])
regimealsacemoselle : type de régime	Entiers dans [99 ; 1 ; 2] ([non applicable ; régime local Alsace Moselle ; complémentaire CAMIEG])
libelle_ape_etab : APE de l'établissement	[Nettoyage courant des bâtiments ; Aide a domicile ; …]
secteur_ent : secteur lourd, moyen ou léger de l'entreprise	[R ; L] ([lourd ; léger])

Pour le modèle de durée, toutes les variables du tableau ci-dessus sont intégrées, et les variables suivantes sont ajoutées.

Variables retenues pour le modèle de durée

Variables numériques	Modalités prises
nbjourpresenceentreprise : nombre de jours de présence en entreprise pour un salarié et son contrat par année	Réels dans [1 ; 365]
tranchesalaire : tranches de salaire mensuel brut	[Moins de 1000 € ; 2000 à 3000 € ; … ; 7000 € et plus]
trancheanciennete : tranches d'ancienneté du salarié dans l'entreprise	[Inférieure à 1 an ; 1 et 2 ans ; ... ; Supérieure à 10 ans]

La durée de présence en entreprise ne peut pas être intégrée dans le modèle de fréquence car elle contribue à la variable cible à prédire : nombre d'arrêts / exposition. Mais elle est sélectionnée pour le modèle de durée.
Les tranches de salaire sont construites d'après le salaire mensuel brut déclaré au moment de l'arrêt.
Les tranches d'ancienneté concernent seulement les salariés sinistrés.

Maintenant que nous avons une base d'apprentissage pour chacune des deux lois, il est nécessaire de séparer la table en deux échantillons : celui qui entraine le modèle et celui qui permet de le valider.

2.2.4.5 Séparation en bases Train et Test

Afin d'obtenir une table qui va servir pour la modélisation, la base est séparée en deux parties Train et Test. Les modèles sont ajustés sur la partie Train et sont validés sur la partie Test selon les performances observées.

Pour ne pas biaiser nos prédictions, il faut qu'un même contrat d'un salarié appartienne soit à l'échantillon Test soit au Train. Un même contrat ne peut pas être dans les deux car nous ne pouvons pas tester notre modèle sur un même contrat que celui qui a servi pour la base d'apprentissage et qui a les mêmes caractéristiques. La base est séparée en fonction des identifiants des contrat des salariés « idsalarie ». Nous créons ainsi une liste qui contient de manière unique les identifiants des contrats qui est ensuite séparée en deux : 80% de la liste sert à constituer le Train et 20% le Test. Ensuite, grâce à la clé identifiant le contrat, les échantillons Train et Test sont obtenus en fusionnant la base d'apprentissage avec la liste des 80% des salariés pour créer le Train et avec les 20% des salariés pour créer le Test.

Cependant, il n'a pas été fait de stratifications par variables pour que les échantillons Train et Test contiennent approximativement la même population. Par précaution, nous vérifions que les deux échantillons ont la même distribution globale et la même répartition sur quelques variables à priori déterminantes pour la tarification : l'âge, le sexe, le statut.

Comme il est possible de le voir sur les statistiques et les graphiques ci-dessous, pour le modèle de fréquence, les échantillons Train et Test ont les mêmes proportions sur les variables étudiées. Ceci peut être expliqué par le volume de la base qui est très important et qui permet, même avec un échantillonnage aléatoire, d'avoir des échantillons Train et Test qui sont similaires dans les statistiques.

Les statistiques effectuées sur le nombre d'arrêts dans les bases Train et Test sont affichées ci-dessous :

Nombre d'arrêts de l'échantillon Train

count	226751.000000
mean	0.201516
std	0.632274
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	16.000000

Nombre d'arrêts de l'échantillon Test

count	56673.000000
mean	0.204736
std	0.625958
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	13.000000

Figure 11 : description statistique du nombre d'arrêts dans le Train et dans le Test

Puis, voici la répartition des années DSN, du sexe et du statut du salarié dans le Train et dans le Test.

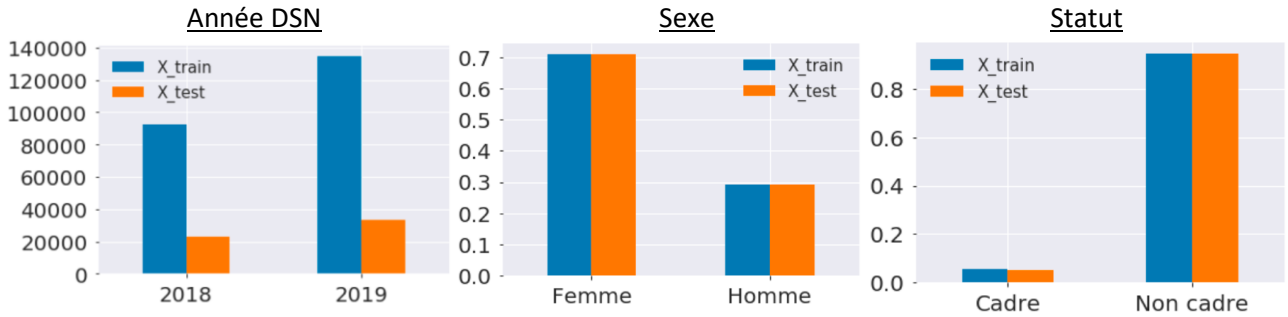


Figure 12 : répartition de l'année DSN, du sexe et du statut du salarié selon les échantillons Train et Test

Et enfin, voici la répartition des âges dans le Train et dans le Test.

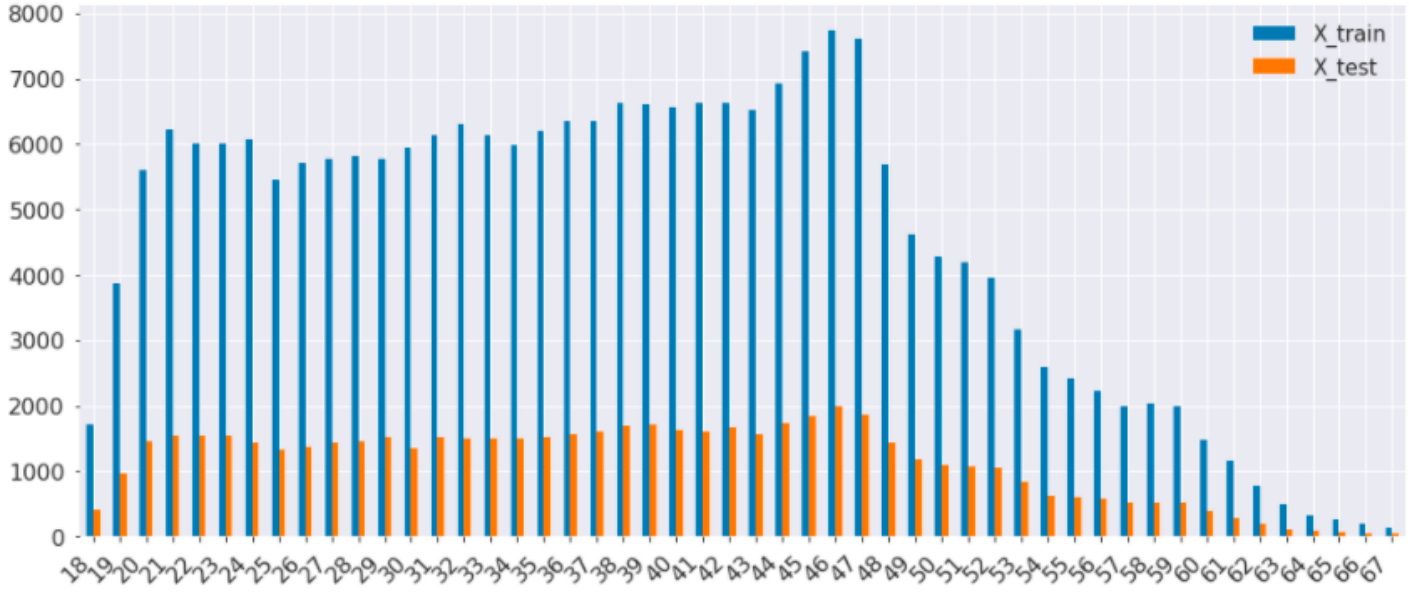


Figure 13 : répartition des âges dans le Train et dans le Test

Nous constatons que les proportions et les statistiques restent quasiment équivalentes dans le Train et dans le Test pour le modèle de fréquence. Pour le modèle de durée, les statistiques ne sont pas présentées ici mais sont aussi stables dans les échantillons Train et Test.

Les bases de données sont à présent exploitables pour effectuer des modélisations sur l'arrêt de travail.
Avant de commencer les travaux de modélisation, des statistiques descriptives sont présentées en guise d'analyse exploratoire pour pouvoir se familiariser avec nos données.

Section 2.3 : les statistiques descriptives

Les statistiques qui sont présentées dans cette partie sont effectuées sur les deux branches assez spécifiques « aide à domicile » et « propreté ». Les autres filtres ainsi que les recatégorisations de variables ne sont pas mis en place pour les statistiques afin d'avoir une vision plus globale.

2.3.1 Statistiques du portefeuille

Le nombre d'entreprises par année est calculé à partir du code SIREN qui est un code attribué à chaque entreprise.

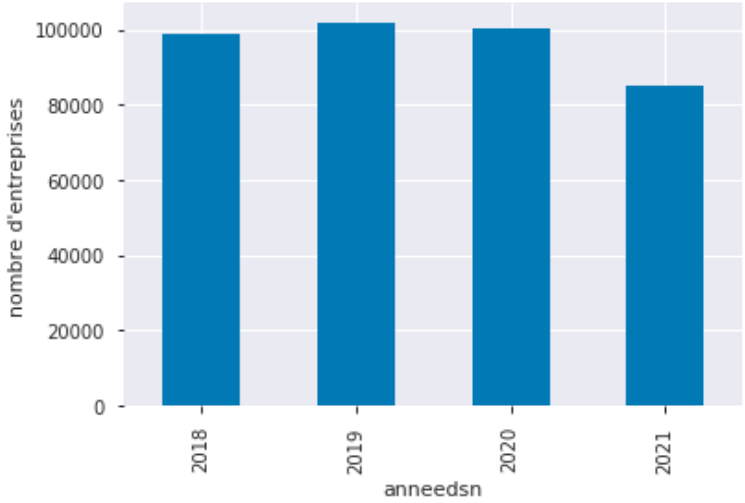


Figure 14 : évolution du nombre d'entreprises par année

Nous constatons que le nombre d'entreprises atteint son pic pour l'année DSN 2019 avec un nombre proche de 100 000. (L'année 2021 n'étant pas finie, il n'est pas possible de se baser sur son nombre d'entreprises actuel.)

Il est intéressant de regarder quels secteurs d'activité (APE) sont les plus présents dans le portefeuille étudié. Sur 708 codes APE présents dans la base, voici les 10 premières activités les plus représentées.

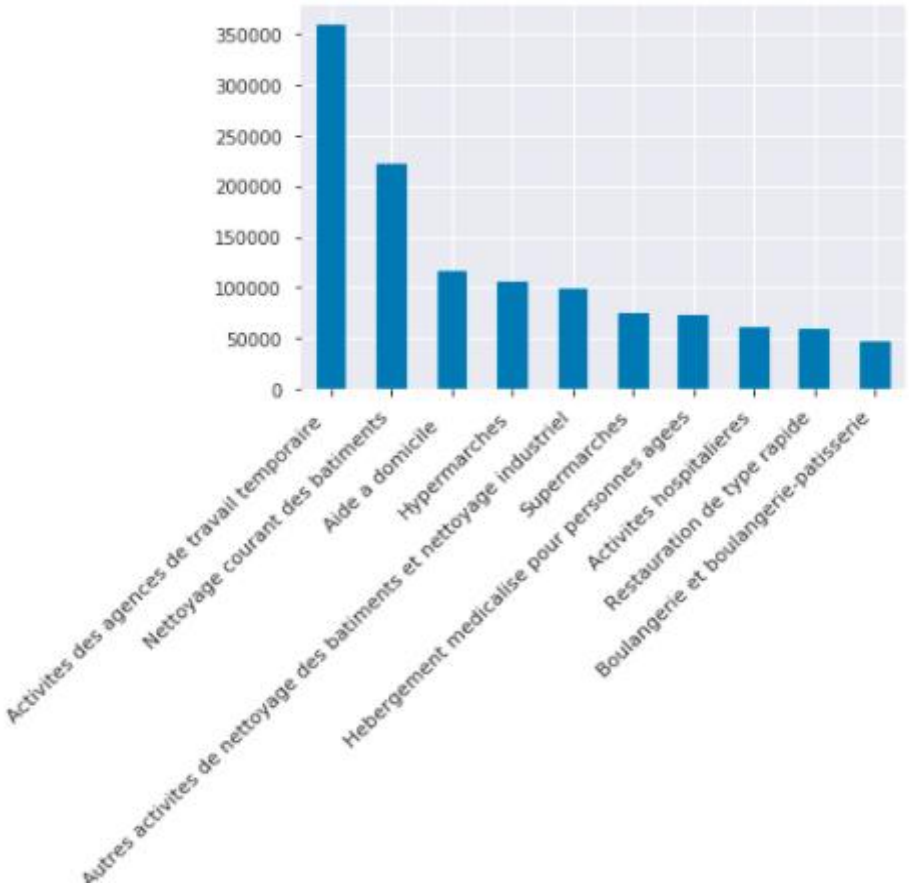


Figure 15 : secteurs d'activité les plus présents dans le portefeuille

Et enfin, voici la répartition des salariés selon leur statut sinistré ou non sinistré. Ce graphique représente la proportion de salariés ayant eu au moins un arrêt de travail en 2018 ou 2019 sur la population totale assurée en 2018 ou 2019. Ceci est important car notre étude repose sur l'absentéisme et nous devons nous assurer qu'il y ait assez de volume pour les sinistres.

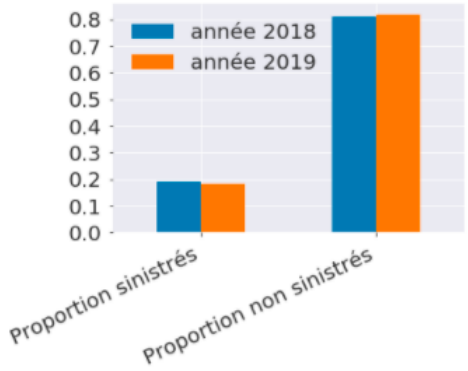


Figure 16 : répartition sinistrés/non sinistrés

Nous observons un écart significatif entre le nombre de personnes ayant eu un arrêt de travail (personnes sinistrées) et celles non sinistrées avec environ 20% des salariés qui sont sinistrés en 2018, ce qui n'est pas négligeable. Ce pourcentage reste assez stable de 2018 à 2019.

Après avoir vu les statistiques globales du portefeuille, il est possible d'analyser plus en détail l'influence des variables sur les arrêts, tant sur leur fréquence, que sur leur durée.

2.3.2 Analyse de l'influence des variables sur le nombre d'arrêts et leurs durées

Clé de lecture :

L'histogramme en jaune-vert représente la répartition des salariés, c'est à dire les pourcentages de salariés pour chaque catégorie (tranche d'âge, sexe…). La courbe rouge indique le niveau de durée d'arrêt et est créée en calculant le pourcentage de jours d'arrêts (somme du nombre de jours d'arrêts dans la catégorie X divisée par la somme totale des durées d'arrêts dans la base) par catégorie. La courbe bleue indique le pourcentage d'arrêts (somme du nombre d'arrêts dans la catégorie X divisée par la somme totale des arrêts dans la base) par catégorie.

Nombre d'arrêts = ———

Durée d'arrêt = ———

Les graphiques suivants montrent les statistiques de 2018. Ceux de 2019 ne sont pas présentés car ils sont quasiment similaires.

2.2.2.1 Répartition des salariés et de leurs arrêts selon l'âge en 2018

Nous distinguons ici la répartition des salariés et des arrêts pour les hommes et pour les femmes. Le premier graphique concerne les hommes, et le graphique du dessous concerne les femmes.

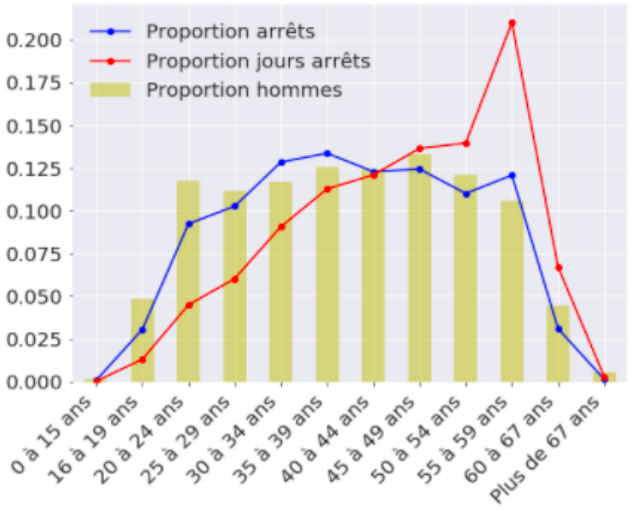


Figure 17 : répartition des hommes et de leurs arrêts selon l'âge en 2018

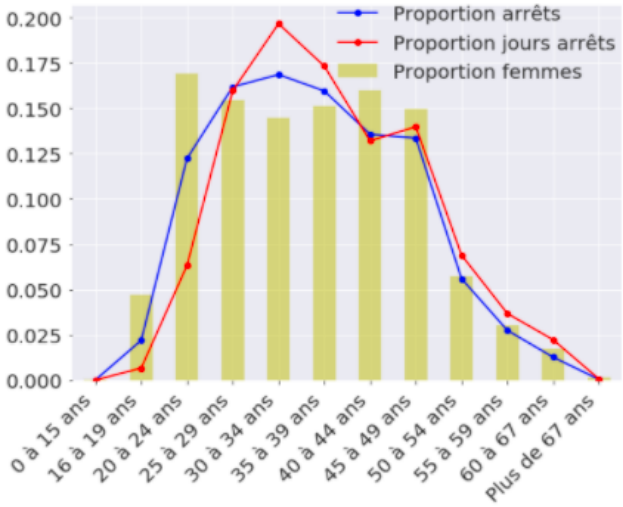


Figure 18 : répartition des femmes et de leurs arrêts selon l'âge en 2018

Clé de lecture :

Si un point rouge se situe au-dessus (resp. en-dessous) de la barre jaune-verte alors l'arrêt est anormalement long (resp. court) pour cette catégorie.
Si un point bleu se situe au-dessus (resp. en-dessous) de la barre jaune-verte alors l'arrêt est anormalement fréquent (resp. rare) pour cette catégorie.
Exemple : les hommes de 50 à 54 ans représentent 12,4% de la population masculine, mais représentent seulement 11% des arrêts observés (ils s'arrêtent donc moins en fréquence). Pour autant, leurs arrêts représentent 14% de la durée totale des arrêts masculins : les arrêts sont donc moins nombreux, mais plus longs que la moyenne.

Nous constatons pour les femmes :

- Un pic pour les durées d'arrêts pour la tranche d'âge de 30 à 39 ans.
  Il peut être interprété par la tendance des grossesses dans ces âges. Cela explique pourquoi les arrêts de travail liés à la maternité sont éliminés car ils peuvent biaiser nos modélisations.

- Des durées d'arrêts plus importantes proportionnellement aux arrêts pour la tranche d'âge de 50 à 67 ans.
- Un faible niveau d'arrêts pour les tranches d'âge de 16 à 24 ans et de 40 à 49 ans par rapport aux autres tranches d'âge.

Pour les hommes, nous constatons :

- Des niveaux d'arrêts plus importants pour les tranches d'âge de 30 à 39 ans et de 55 à 59 ans. Ces constats peuvent être expliqués par les arrêts paternité (30 à 39 ans) et la présence de risques de maladies cardiovasculaires (55 à 59 ans).
- Des longues durées d'arrêts pour la tranche d'âge de 50 à 67 ans par rapport aux autres tranches.
- Moins d'arrêts et surtout des durées moins longues pour les salariés qui ont entre 16 et 29 ans que pour les autres.

Les hommes et les femmes qui ont entre 30 et 39 ans ont donc généralement plus d'arrêts que les autres, mais les arrêts sont moins longs pour les hommes que pour les femmes dans ces tranches d'âge. Pour les hommes, les arrêts sont particulièrement longs après 50 ans.

2.2.2.2 Répartition des salariés et des arrêts selon le sexe en 2018

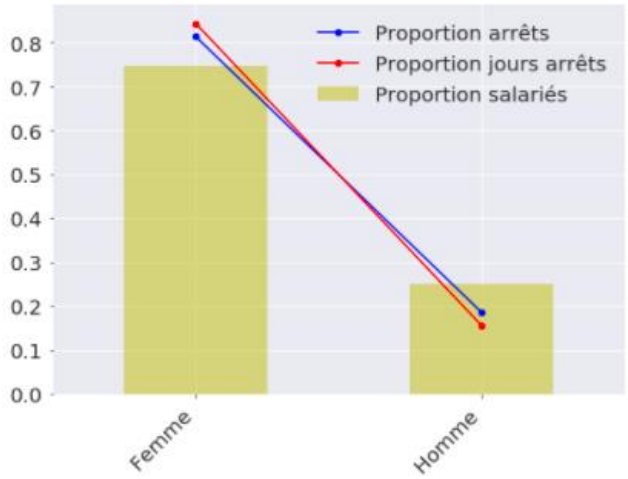


Figure 19 : répartition des salariés et de leurs arrêts selon le sexe en 2018

Nous constatons un pourcentage plus important de femmes (75%) que d'hommes (25%) dans le portefeuille. Cette observation est due au choix du périmètre des branches « aide à domicile » et « propreté » dans lesquelles les femmes sont bien plus présentes que les hommes. Aussi, nous observons que les arrêts sont plus nombreux et plus longs pour les femmes. Ce résultat n'a rien de surprenant car les femmes ont généralement plus d'arrêts avec les grossesses, et les arrêts maternité sont assez longs.

2.2.2.3 Répartition des salariés et de leurs arrêts selon le salaire mensuel en 2018

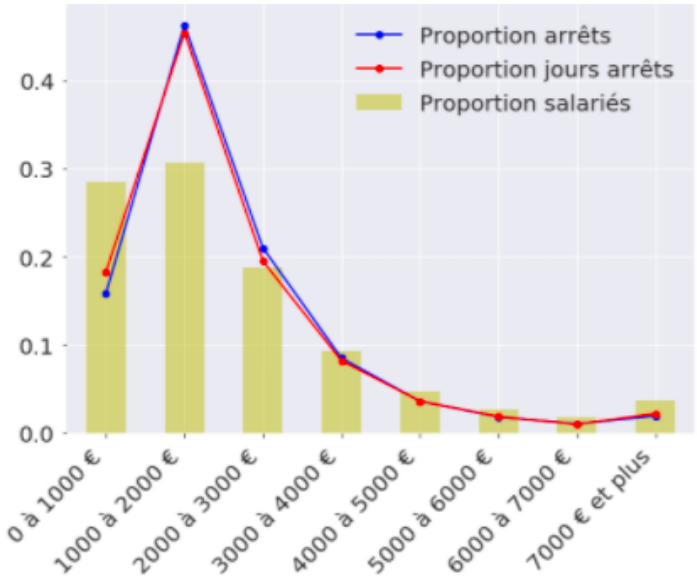


Figure 20 : répartition des salariés et de leurs arrêts selon le salaire mensuel en 2018

Nous constatons que 60% des salariés du portefeuille touchent moins de 2000 € brut par mois. Cela s'explique par le fait que les salariés sont majoritairement en temps partiel dans les branches étudiées.

Nous notons également un pic d'arrêts pour les salariés ayant des rémunérations allant de 1000 à 2000 €. Pour les salaires extrêmes, ceux de plus de 7000 € et ceux de moins de 1000 €, nous remarquons un faible nombre d'arrêts par rapport aux autres tranches de salaire.

2.2.2.4 Répartition des sinistrés et de leurs arrêts selon la région d'arrêt en 2018

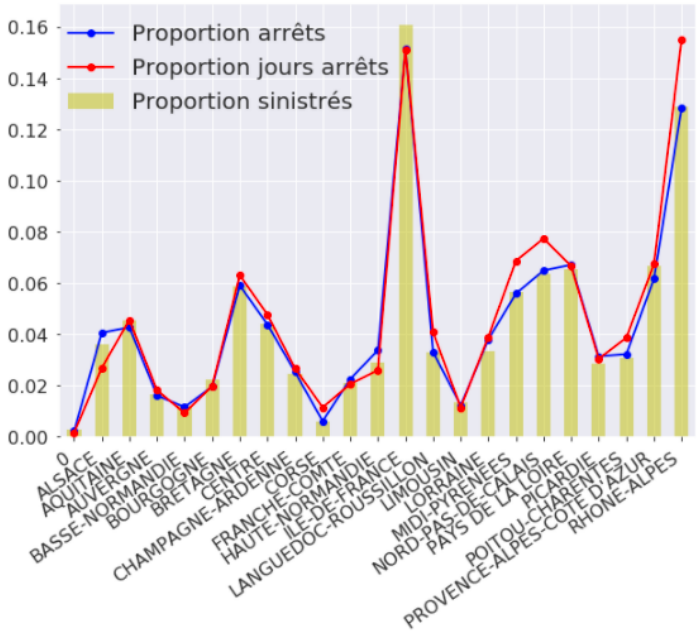


Figure 21 : répartition des sinistrés et de leurs arrêts selon la région d'arrêt en 2018

Dans notre portefeuille étudié, les trois régions les plus présentes sont l'Ile de France (16%), la région Rhône-Alpes (13%) et la région Provence-Alpes-Côte D'Azur (7%).

Proportionnellement au nombre de salariés, les régions avec le plus d'arrêts sont l'Alsace, la Haute-Normandie et les Midi-Pyrénées et celles possédant le moins d'arrêts sont l'Ile de France et Provence-Alpes-Côte d'Azur.

Concernant les durées d'arrêts moyennes, elles sont particulièrement importantes en Rhône-Alpes, Midi-Pyrénées et Nord-Pas-de-Calais et plus basses en Alsace.

2.2.2.5 Répartition des sinistrés et de leurs arrêts selon le motif d'arrêt en 2018

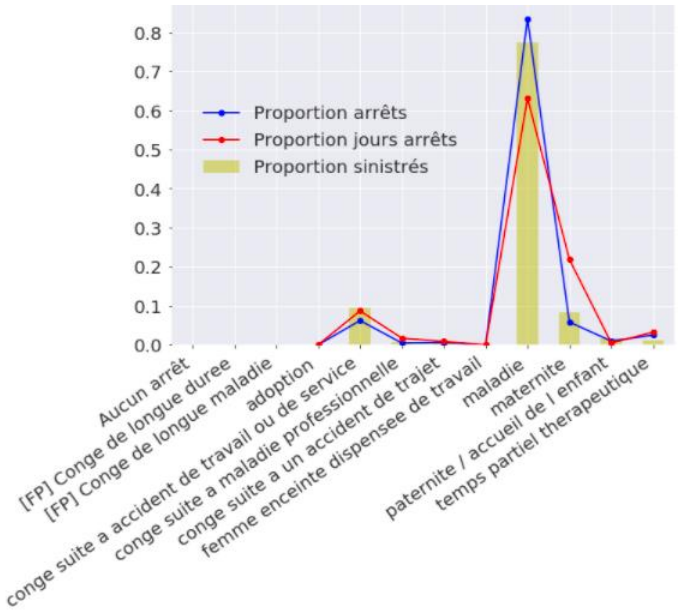


Figure 22 : répartition des sinistrés et de leurs arrêts selon le motif d'arrêt en 2018

Les motifs d'arrêts les plus fréquents sont ceux liés à la maladie (79%), aux accidents du travail (10%), puis à la maternité (8%).

Nous constatons un pic d'arrêts pour le motif « maladie » avec des durées d'arrêts assez faibles comparées aux arrêts particulièrement longs pour la maternité. Ce dernier constat est tout à fait logique car la maternité impose généralement une durée d'arrêt importante. Afin de ne pas biaiser notre étude, ces arrêts assez spécifiques ne sont pas intégrés dans notre modélisation de l'arrêt de travail.

2.2.2.6 Répartition des salariés et de leurs arrêts selon le statut en 2018

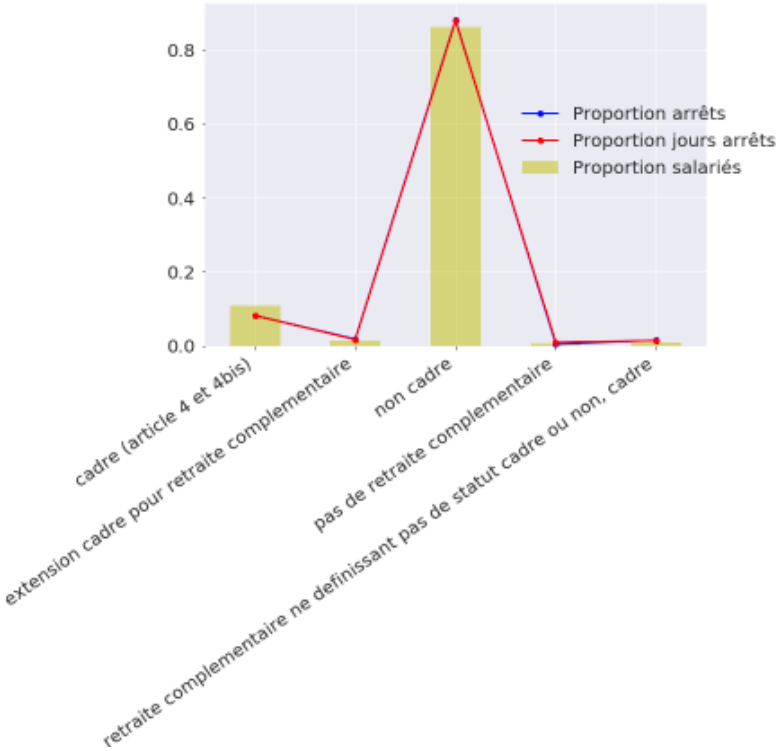


Figure 23 : répartition des salariés et de leurs arrêts selon le statut en 2018

Il y a 90% de non-cadres dans notre portefeuille, contre 10% de cadres. Ce dernier statut est plus rare étant donné notre périmètre de branches choisi.

Nous observons que les arrêts sont légèrement plus importants pour les non-cadres que les cadres. En effet, ce résultat peut être interprété par la situation des cadres qui est souvent plus confortable que celle des non - cadres.

2.2.2.7 Répartition des salariés et de leurs arrêts selon la taille de l'établissement en 2018

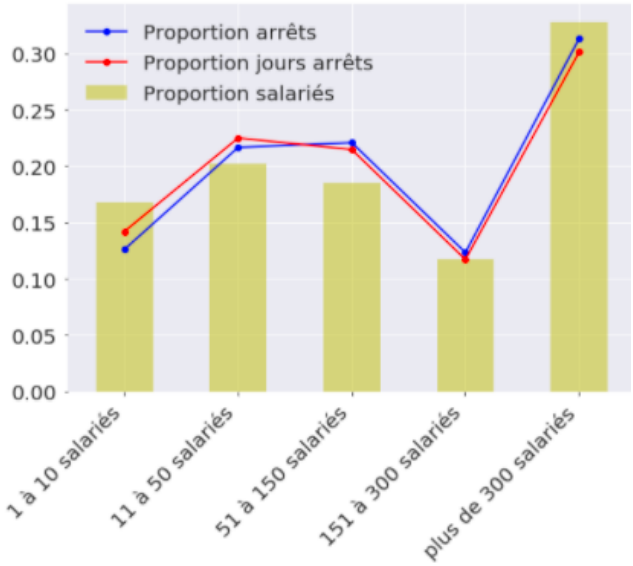


Figure 24 : répartition des salariés et de leurs arrêts selon la taille de l'établissement en 2018

Les entreprises les plus présentes dans notre portefeuille sont constituées de plus de 300 salariés avec un taux de 30%.

Le niveau d'arrêts est plus faible pour les très petites entreprises et pour les très grandes. Pour les entreprises de 1 à 10 salariés et celles de plus de 300 salariés, il y a moins d'arrêts que pour les autres. Pour les entreprises de 51 à 150 salariés, c'est-à-dire pour les entreprises de taille moyenne, il y a plus d'arrêts.

2.2.2.8 Répartition des salariés et de leurs arrêts selon la CSP en 2018

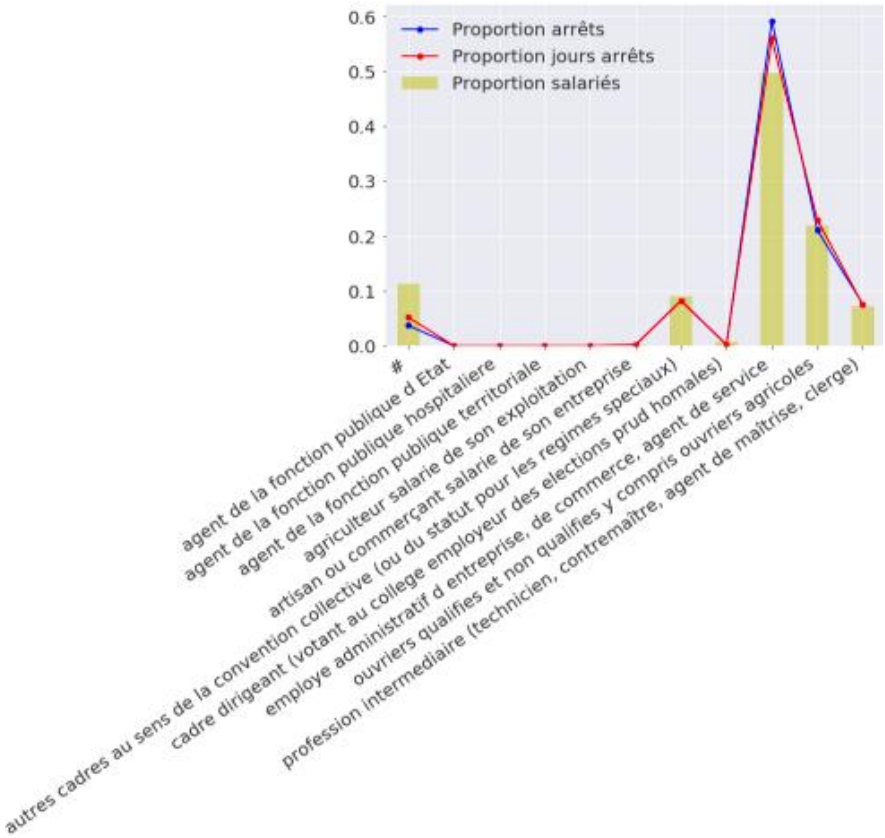


Figure 25 : répartition des salariés et de leurs arrêts selon la CSP en 2018

La moitié du portefeuille est constituée d'employés administratifs d'entreprise, de commerce et d'agents de service. L'autre partie est constituée de 22% d'ouvriers qualifiés et non qualifiés y compris les ouvriers agricoles, de 9% de cadres au sens de la convention collective, de 8% de professions intermédiaires et de 11% de salariés avec une CSP inconnue.

Cumulant 60% des arrêts pour seulement 50% de la population assurée, les employés administratifs d'entreprise, de commerce et les agents de service sont les CSP les plus touchées par l'arrêt de travail.

2.2.2.9 Répartition des sinistrés et de leurs arrêts selon les mois de survenance en 2018

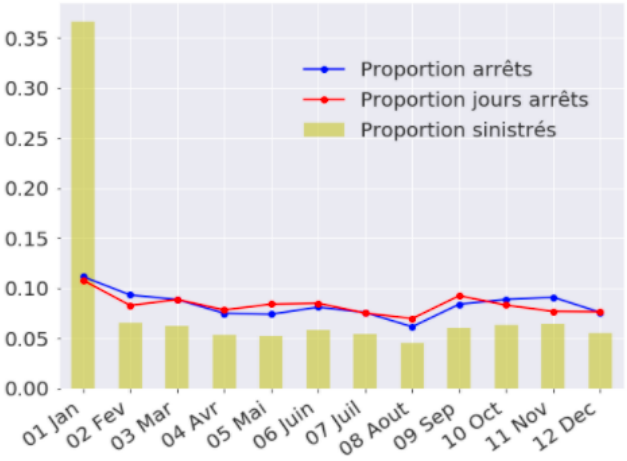
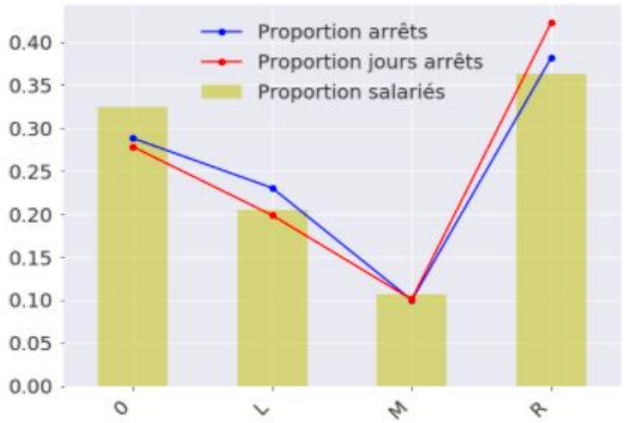


Figure 26 : répartition des sinistrés et de leurs arrêts selon les mois de survenance en 2018

Il y a plus de personnes qui s'arrêtent en janvier dans notre portefeuille avec un taux d'environ 37%. Mais, ces sinistrés ont moins d'arrêts que ceux qui ont des sinistres durant les autres mois. Et nous remarquons que de manière générale, durant la période estivale (mai-septembre), les arrêts sont plus longs que durant la période hivernale. L'été est effectivement une période plus profitable qui peut influencer la durée des arrêts.

2.2.2.10 Répartition des salariés et de leurs arrêts selon le secteur en 2018



L : léger

M : moyen

R : lourd

O : non classé

Figure 27 : répartition des salariés et de leurs arrêts selon le secteur en 2018

Nous observons qu'environ 36% des salariés sont dans un secteur d'activité lourd contre 21 % dans un secteur léger. Cela s'explique par les branches « propreté » et « aide à domicile » choisies qui ont plus tendance à être liées à des entreprises à secteur lourd en raison du type d'activité à risques psychiques et physiologiques importants.

Les arrêts sont plus courts dans les secteurs L (léger). A l'inverse, il y a des arrêts plus longs dans le secteur R (lourd). Ceci se justifie par le fait que dans les secteurs lourds comme les activités d'entretien où les individus sont plus à risque, les accidents peuvent être plus graves que dans d'autres secteurs comme l'informatique par exemple. Un individu qui a une fracture du pied pourra continuer à exercer l'informatique, tandis qu'un agent d'entretien sera dans l'impossibilité et l'arrêt sera donc plus long.

2.2.2.11 Répartition des salariés et de leurs arrêts selon la modalité de l'exercice en 2018

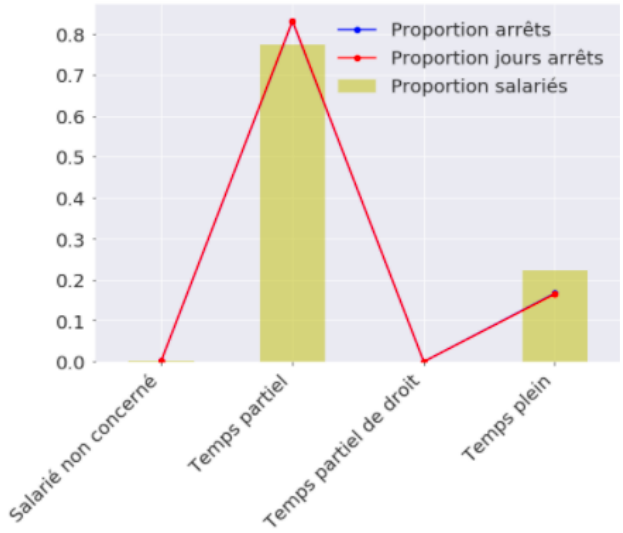


Figure 28 : répartition des salariés et de leurs arrêts selon la modalité de l'exercice en 2018

Il y a 80% des salariés en temps partiel en 2018, et 20% des salariés en temps plein. Cela s'explique encore une fois par les branches choisies dans lesquelles, les salariés sont majoritairement en temps partiel. Nous constatons aussi qu'il y a plus d'arrêts pour les salariés en temps partiel que pour les salariés en temps plein, en raison de leur situation qui est plus précaire.

2.2.2.12 Ratios jours arrêtés / jours de présence en entreprise selon les tranches d'âge en 2018 et 2019

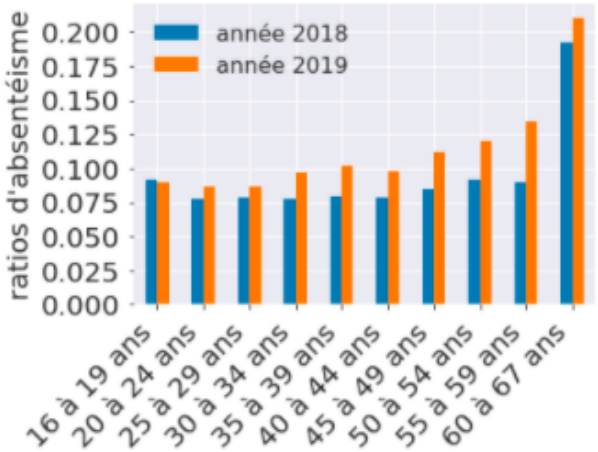


Figure 29 : ratios jours arrêtés / jours de présence en entreprise selon les tranches d'âge en 2018 et 2019

Clé de lecture :

Les ratios d'absentéisme (durée d'arrêt moyenne/ durée de présence en entreprise moyenne par catégorie) sont plus importants pour les tranches 55 à 67 ans avec un taux allant de 10% à 13%. Cela signifie que pour cette tranche d'âge, les salariés sont plus absents proportionnellement à leur présence en entreprise. Pour 100 jours de présence en jours calendaires, ils ont 10 à 13 jours arrêtés. Pour les moins de 30 ans, ce taux est environ égal à la moitié.

Nous constatons que les ratios d'absentéisme augmentent de 2018 à 2019. Cela confirme la dégradation des taux d'absentéisme vue dans le baromètre Ayming.
Concernant l'évolution du taux en 2019, nous constatons une baisse du taux jusqu'à 29 ans, puis une augmentation globale jusqu'à 67 ans.

Après avoir constaté les différents effets des variables sur les arrêts, il est possible de comparer la répartition des sinistrés et des non sinistrés au sein de différentes catégories pour approfondir cette analyse avec une approche différente.

2.3.3 Comparaison entre les sinistrés et les non sinistrés

Un salarié est considéré comme sinistré lorsqu'il tombe en arrêt au moins une fois dans l'année.

Clé de lecture :

Les histogrammes oranges représentent les pourcentages de salariés qui sont non sinistrés dans les différentes catégories et les histogrammes bleus nous montrent les proportions des salariés sinistrés dans celles-ci.

2.2.3.1 Répartition des salariés par tranches d'âge en 2018

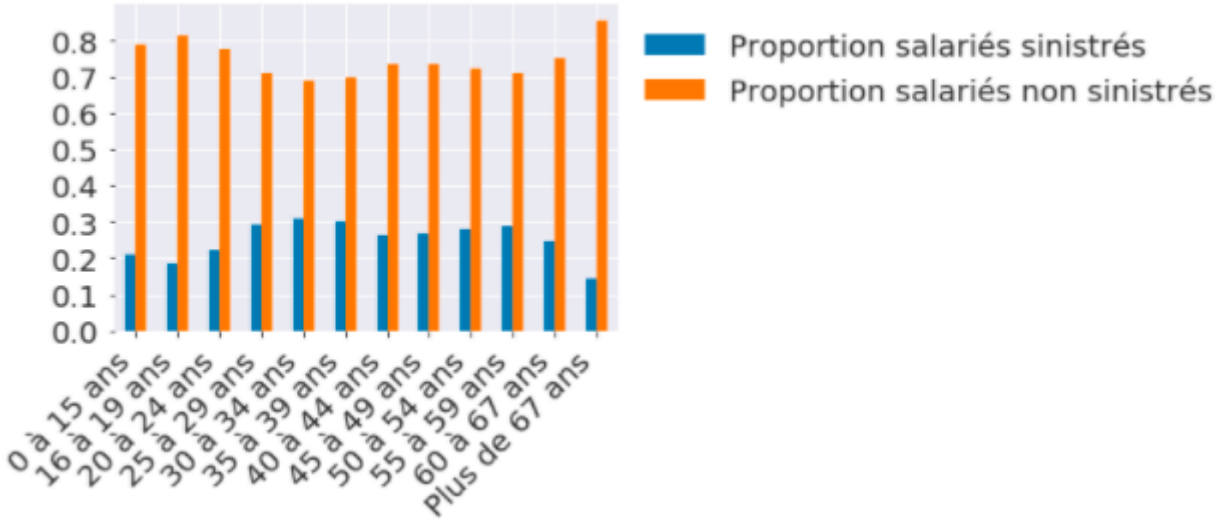


Figure 30 : répartition des salariés par tranche d'âge en 2018

Les sinistrés sont répartis de manière assez homogène sur les différentes tranches d'âge. Mais, nous remarquons tout de même une hausse des sinistrés entre 25 et 39 ans avec un taux maximum de 30% des salariés qui sont sinistrés. Celle-ci est due aux tendances des arrêts maternité et paternité dans cette tranche d'âge.

2.2.3.2 Répartition des salariés par tranche de salaire en 2018

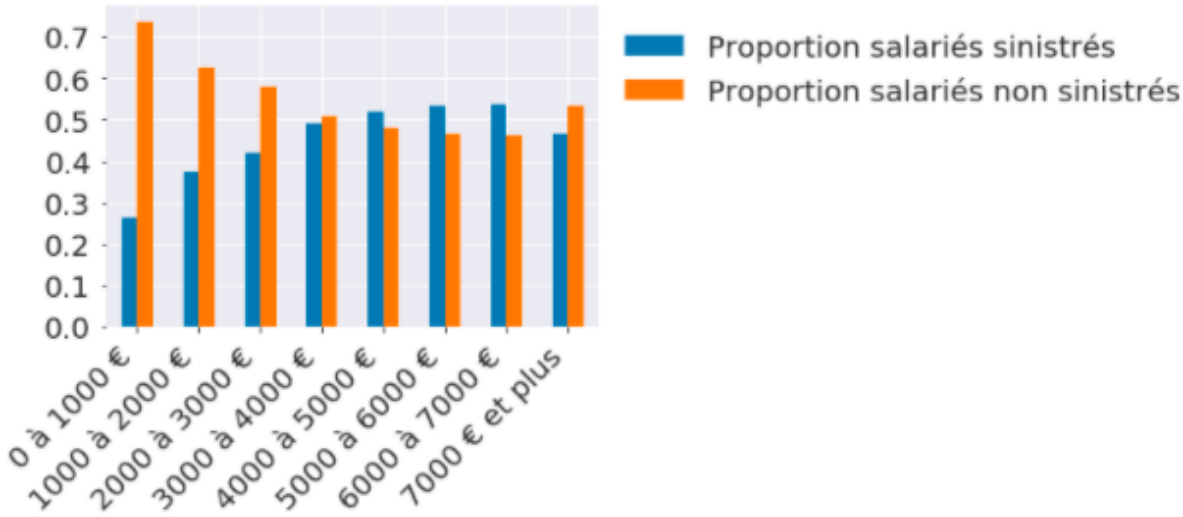


Figure 31 : répartition des salariés par tranche de salaire en 2018

Les sinistrés sont majoritairement présents par rapport aux non sinistrés sur les salaires importants allant de 3000 euros à 7000 euros atteignant un taux de 55% des salariés qui sont sinistrés contre ceux non sinistrés (pour ceux qui gagnent entre 5000 à 6000 €), ce qui représente plus de la majorité. Nous remarquons que pour les faibles salaires, ce sont les salariés non sinistrés qui sont le plus représentatifs avec un taux de 72% des salariés qui gagnent moins de 1000 euros qui n'ont pas eu d'arrêts contre 28 % des non sinistrés.

Les CSP qui ont des situations plus stables permettent l'arrêt alors que les personnes moins aisées peuvent moins s'autoriser cela. Aussi, le salaire est corrélé à l'âge et nous avons vu que les personnes âgées s'arrêtent plus (les hommes de 55 à 60 ans), ce qui justifie les précédents constats.

2.2.3.3 Répartition des salariés par taille d'établissement en 2018

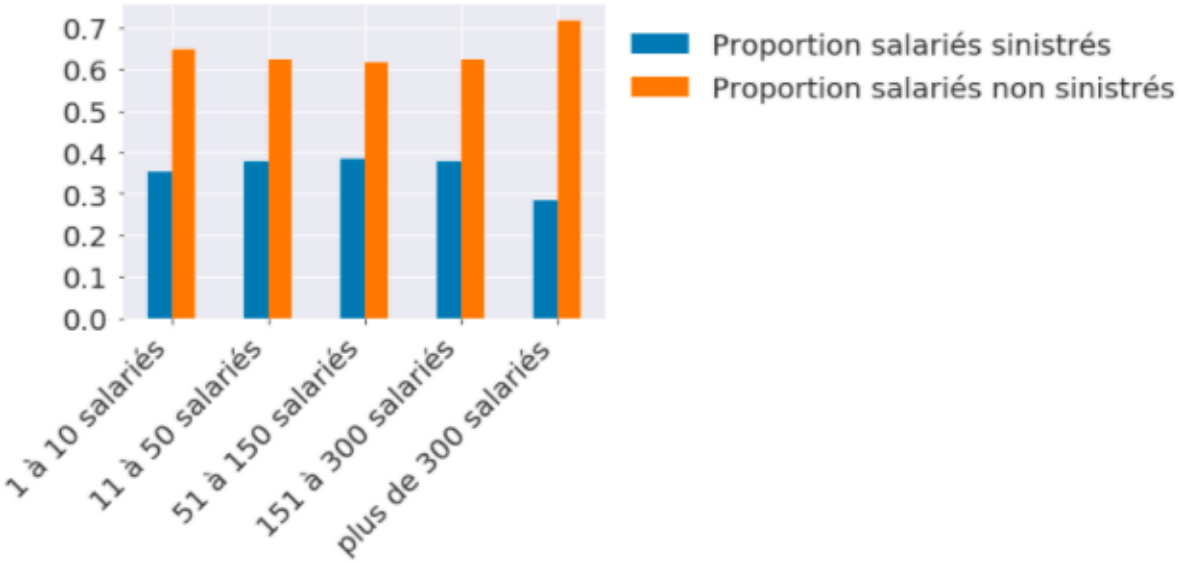


Figure 32 : répartition des salariés par taille d'établissement en 2018

Les salariés sinistrés sont un peu moins représentés dans les entreprises de plus de 300 salariés dans lesquelles environ 30% des salariés s'arrêtent au moins une fois par an alors que cette statistique s'élève à 40% pour les entreprises de 11 à 300 salariés.

2.2.3.3 Répartition des salariés selon la nature des contrats et le statut du salarié en 2018

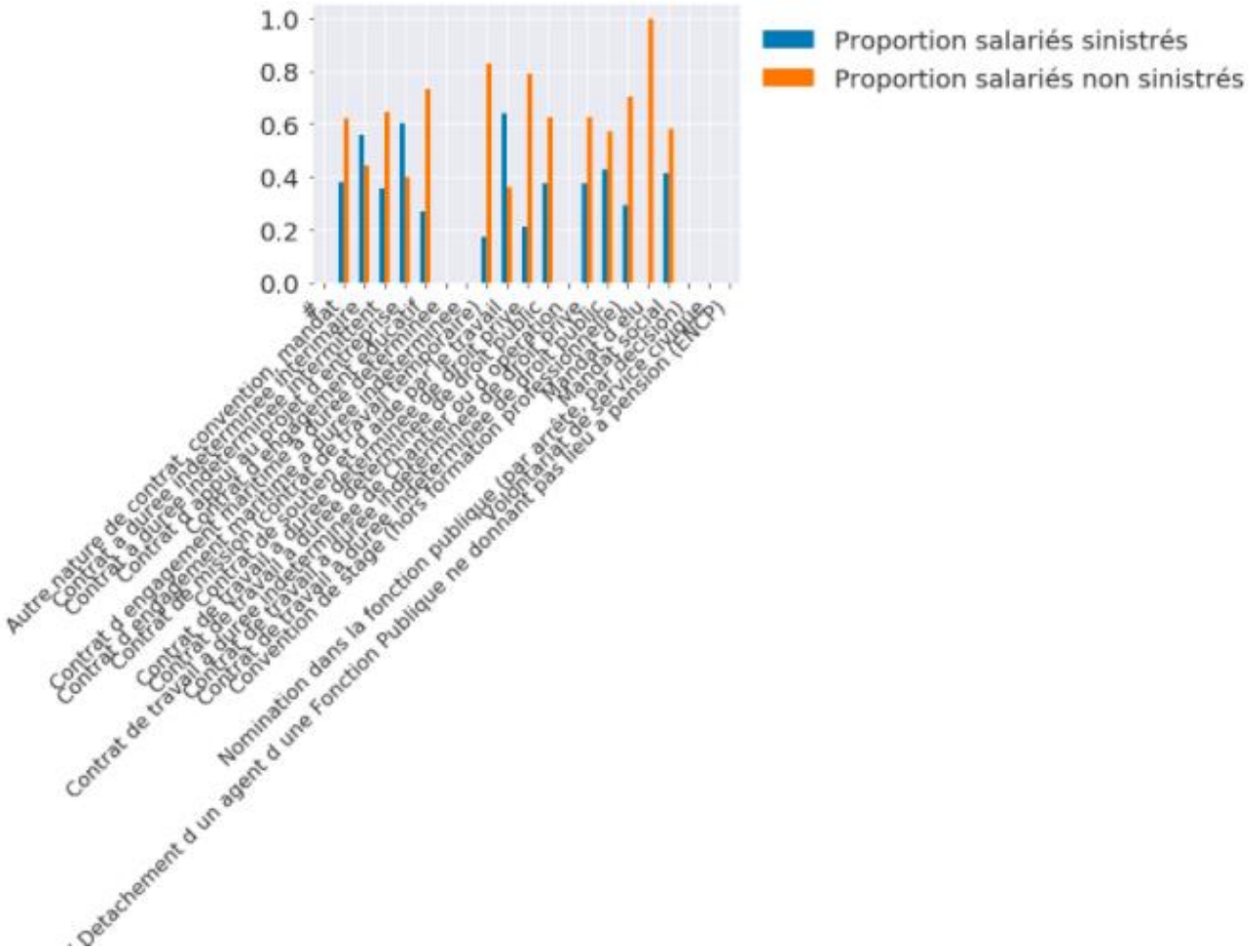


Figure 33 : répartition des salariés selon la nature des contrats et le statut du salarié en 2018

Les sinistrés sont majoritairement présents pour les contrats de soutien et d'aide par le travail, pour les contrats d'appui au projet d'entreprise et pour les contrats à durée intérimaire avec des pourcentages de salariés sinistrés tournants autour de 60% pour chacun de ces types de contrat. Pour les mandats d'élus, les contrats de mission et les contrats de travail de durée indéterminée de droit privé, il y a peu de sinistrés, le taux restant en dessous du seuil de 20%.

2.2.3.4 Distance moyenne (en hm) du domicile au travail par tranche d'âge pour les sinistrés et les non sinistrés en 2018

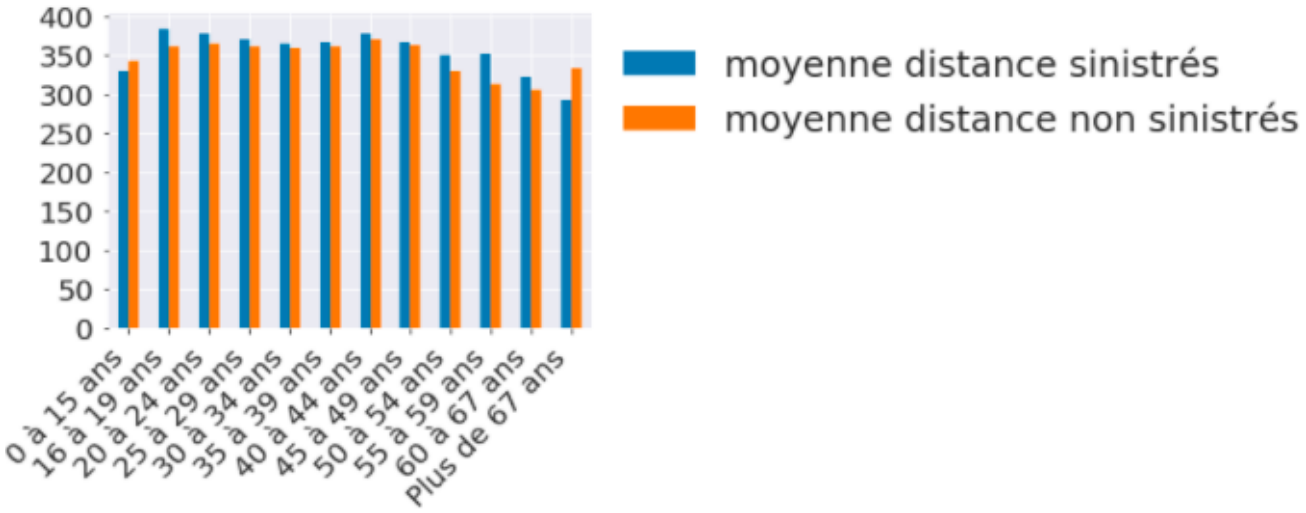


Figure 34 : distance moyenne (en hm) par tranche d'âge pour les sinistrés et les non sinistrés en 2018

Globalement, à part pour les moins de 16 ans et les plus de 67 ans, la distance moyenne du travail à l'entreprise est plus importante pour les sinistrés que pour les non sinistrés. La distance du travail au domicile joue donc défavorablement dans le niveau d'arrêts.

Maintenant que l'influence des différentes variables sur les arrêts et leur durée a été étudiée, les corrélations entre les variables et leur importance pour nos modélisations à venir sont présentées.

Section 2.4 : étude de corrélations

Dans cette section, les corrélations entre les variables sont simplement analysées pour avoir une vision à priori. Dans les paragraphes 2.4.2 et 2.4.3, nous allons plus loin en étudiant les contributions de ces variables dans la variabilité de la base de données.

2.4.1 Corrélations : variables qualitatives et quantitatives

Les corrélations entre les variables quantitatives sont établies grâce aux coefficients de Spearman tandis que les corrélations entre les variables qualitatives sont calculées grâce au V de Cramer.

Harold Cramer instaure le V de Cramer en 1946 afin de mesurer le lien entre deux variables qualitatives qui ont plusieurs modalités, à l'aide de la statistique χ^2 de Pearson. C'est la méthode qui est la plus utilisée par les statisticiens pour l'étude de corrélations entre des variables qualitatives.
Soit X_1 et X_2, deux variables qualitatives avec respectivement K_1 et K_2 modalités et n observations. Voici la formule clé permettant de calculer la corrélation entre X_1 et X_2 :

$$\mathrm{V} = \sqrt{\frac{\chi^2}{n * \min\,(K_1 - 1; K_2 - 1)}}$$

La statistique du χ^2 pour un échantillon de taille n avec par exemple deux variables A et B et $n_{i,j}$ le nombre de fois que les valeurs (A_i , B_i) sont observées est la suivante :

$$\sum_{i,j} \frac{(n_{i,j} - \frac{n_i\, n_j}{n})^2}{\frac{n_i\, n_j}{n}}$$

Le terme du dénominateur est en fait la valeur maximale potentielle de χ^2 et $|\mathrm{V}|$ varie donc entre 0 et 1.

Le coefficient de Spearman est un indice compris entre -1 et 1 qui permet de mesurer l'intensité entre deux variables quantitatives continues ou ordinales. Le principe consiste à utiliser le coefficient de Bravais-Pearson sur les rangs et non sur les observations. En fait, pour chaque observation, le score obtenu est remplacé par son rang. Lorsque le coefficient est égal à 0, cela signifie que les deux variables n'ont pas de lien. Voici la formule permettant d'obtenir le coefficient :

$$p = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x}\, \sigma_{rg_y}}$$

Avec les variables X_i de rang rg_{X_i} et Y_i de rang rg_{Y_i}, leurs écarts-types σ_{rg_x} et σ_{rg_y} respectivement, et leur covariance $cov(rg_x, rg_y)$.

Les corrélations entre les variables sont maintenant étudiées avec trois objectifs. Le premier est d'avoir une première analyse du lien entre les variables afin de mieux comprendre le contexte de notre étude et de mieux nous préparer à la modélisation (analyse exploratoire). Le deuxième objectif est d'éliminer certaines variables redondantes qui répètent l'information et le troisième est de cibler les variables qui ont un lien fort avec a fréquence d'arrêt et la durée de l'arrêt.

Afin d'analyser le lien entre les différentes variables, une matrice de corrélations de Spearman est construite pour les variables quantitatives. Une autre matrice de corrélations est construite à l'aide du V de Cramer pour les variables qualitatives. Les relations entre les variables qualitatives et quantitatives ne sont pas étudiées car les informations quantitatives sont reprises pour la plupart sous forme qualitative (exemple : ancienneté en tranche d'ancienneté). Il est aussi rappelé que cette étude de corrélations est une étude à priori. Nous n'éprouvons donc pas le besoin de regarder les corrélations de manière exhaustive, mais plutôt les corrélations principales.

L'étude est faite sur la population des sinistrés.

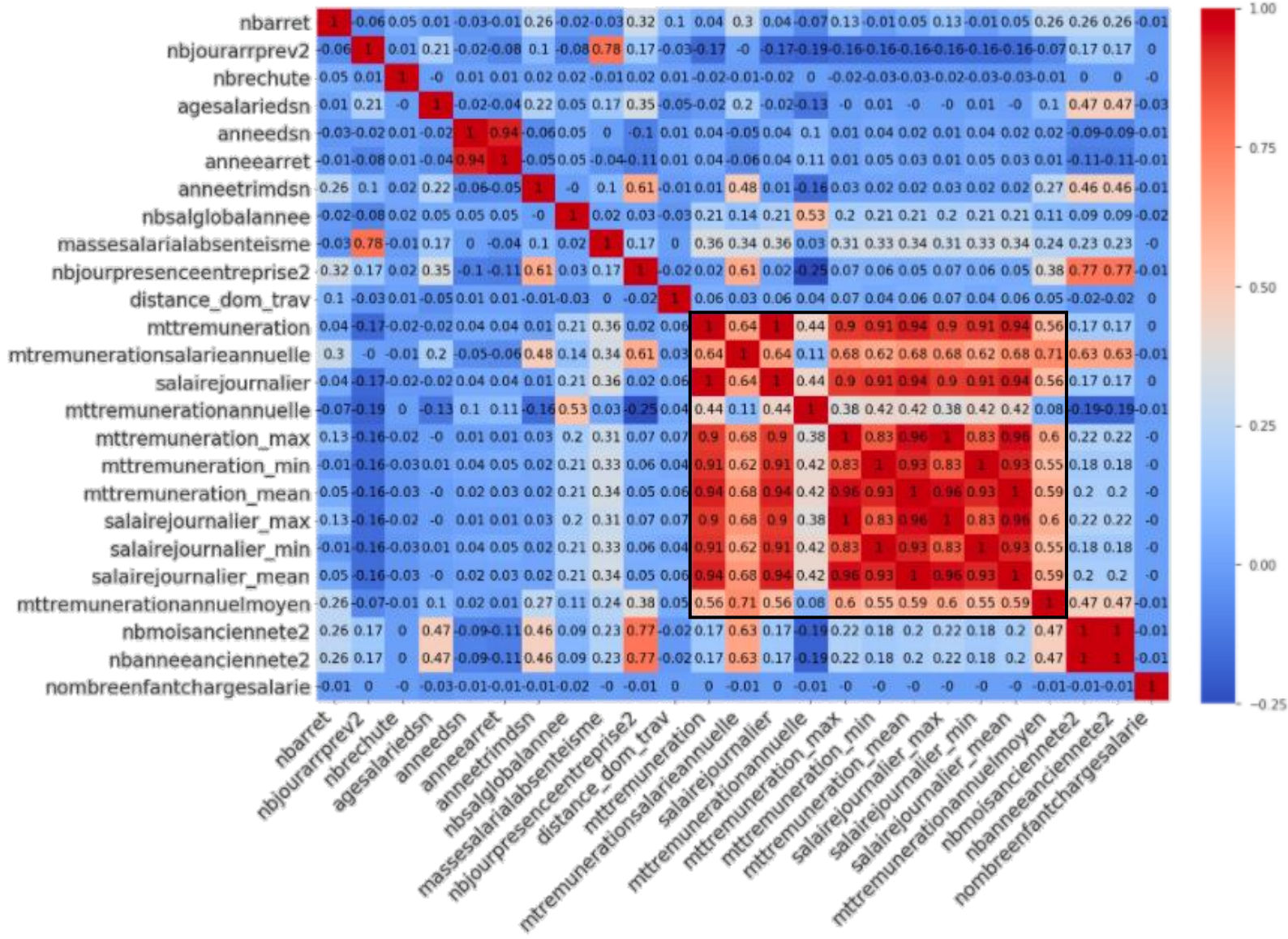


Figure 35 : matrice de corrélations de Spearman[22]

Voici ci-dessous les constats qu'il est possible de faire sur cette matrice.

- Nous remarquons en premier lieu la forte dépendance entre toutes les variables qui contiennent l'information sur le revenu (corrélations encadrées en noir), ce qui est évident.

- Les variables qui sont liées au nombre d'arrêts sont les suivantes : « anneetrimdsn » (0,26), « nbjourpresenceentreprise » (0,32), « mtremunerationsalarieannuelle » (0,3), « mtremunerationannuelmoyen » (0,26) et « nbmois(/annee)anciennete » (0.26).
  En effet, le nombre d'arrêts peut dépendre des trimestres plus ou moins à risques. Mais aussi, la présence et l'ancienneté dans l'entreprise ont une influence sur le nombre d'arrêts. Plus un salarié est présent, plus il est sujet aux arrêts. La situation de confort du salarié a aussi un impact sur ses arrêts et nous comprenons donc que son salaire peut avoir un effet sur son comportement en entreprise.

[22] Le lexique des variables se trouve en **annexe 4**.

- Les variables qui sont liées à la durée de l'arrêt sont : « agesalariedsn » (0,21), « massesalarialabsentesime » (0,78), « nbjourpresenceentreprise » (0,17), les montants de rémunération (coefficients allant jusqu'à - 0,19) et l'ancienneté (0,17).
  Les personnes qui sont plus âgées peuvent avoir des complications dans leurs arrêts (maladie par exemple) et ceux-ci risquent de durer plus longtemps. Un autre exemple concerne les femmes aux alentours de 35 ans qui peuvent avoir des arrêts maternité assez longs comme vu dans les statistiques. Aussi, la masse salariale d'absentéisme est construite à partir des durées d'arrêts, la forte corrélation est donc normale. Les autres variables sont logiquement liées à la durée de l'arrêt pour les mêmes raisons qu'au tiret précédent.

- L'âge du salarié est particulièrement lié à sa durée de présence en entreprise (0, 35) et à son ancienneté (0 ,47). Ces constats sont tout à fait logiques, un salarié ne peut pas être ancien dans l'entreprise lorsqu'il est junior.

- L'année de déclaration de l'arrêt est très liée à son année de survenance (0,94). Cela est rassurant car la déclaration de l'arrêt est la plupart du temps dans la même année que l'arrêt.

- Le trimestre de déclaration de l'arrêt (« anneetrimdsn ») est corrélé à la présence en entreprise (0,61), au « mtremunerationsalarieannuelle » (0,48) et à l'ancienneté (0,46).

- Le nombre de salariés (« nbsalglobalanneee ») est lié au « mtremunerationannuel » (0,53). Ceci paraît normal car, par construction, le montant de rémunération annuel est la somme des salaires de l'entreprise par année.

- Le temps de présence en entreprise est lié au salaire annuel du salarié avec un coefficient de 0,61 et avec l'ancienneté (0,77). L'ancienneté est aussi très corrélée au salaire avec des coefficients de 0,63 pour le montant de rémunération annuel du salarié et 0,47 pour le montant de rémunération mensuel moyen par salarié. Plus une personne est ancienne dans l'entreprise, plus son salaire a des chances d'être augmenté.

Les corrélations entre les variables qualitatives sont maintenant étudiées. Voici la matrice de Cramer.

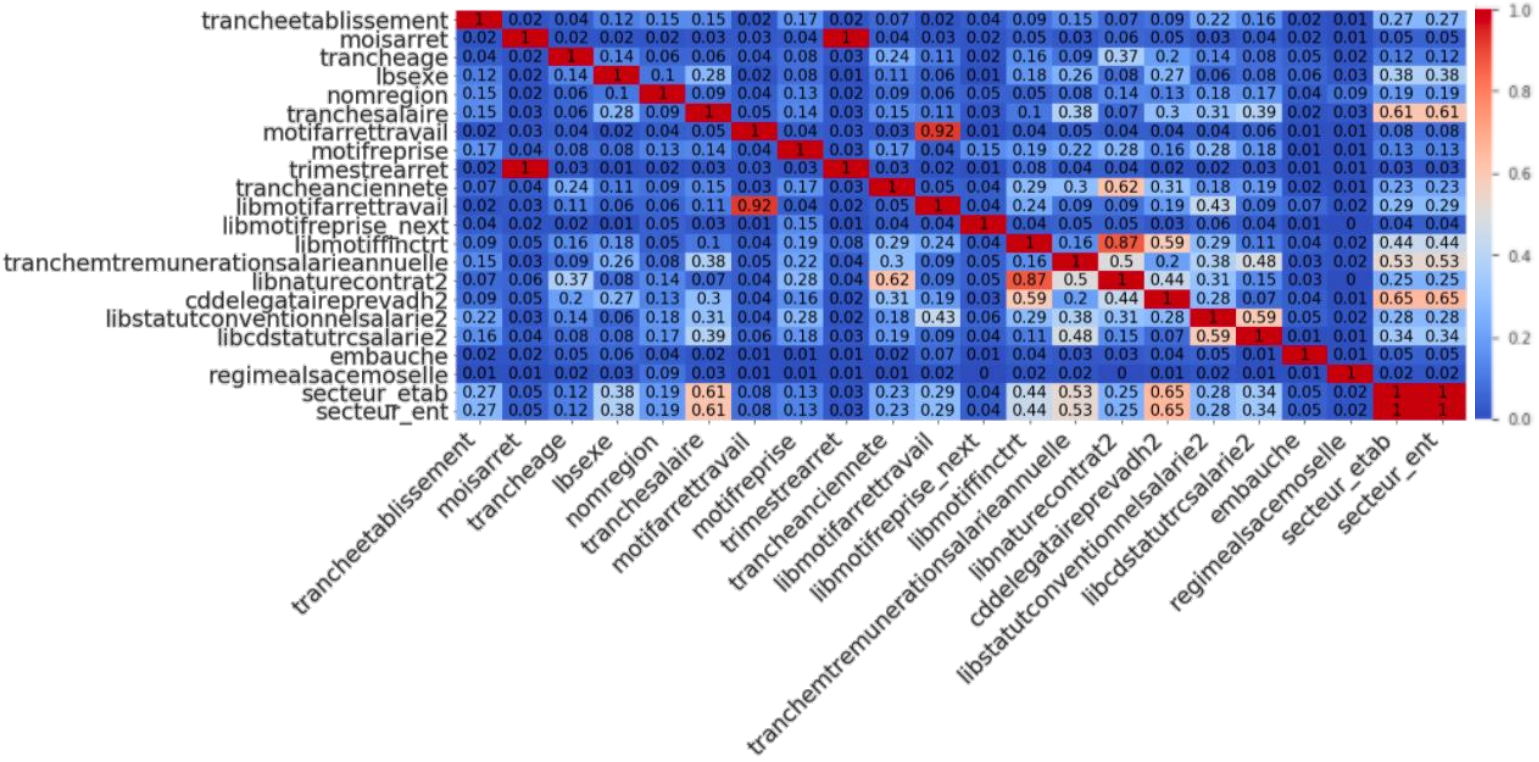


Figure 36 : matrice de corrélations de Cramer[23]

Les corrélations entre les variables repérables dans la matrice sont détaillées ici.

- La taille de l'établissement (« trancheetablissement ») est liée à la région et aux tranches de salaires avec des coefficients égaux à 0,15, au statut conventionnel du salarié avec un indice de 0,22, au motif de reprise (0,17) puis aux secteurs lourds, moyens ou légers de l'établissement et de l'entreprise (0,27).
Cela nous paraît tout à fait logique que l'établissement selon qu'il soit de taille importante ou non soit en lien avec le secteur de l'établissement lourd, moyen ou léger. La taille de l'établissement dépend de son activité et donc de son classement dans les secteurs plus ou moins risqués. De la même manière, ce n'est pas surprenant que la taille de l'établissement est en lien avec les salaires et avec le statut des salariés. Les grandes entreprises sont plus cotées et les salariés sont généralement mieux payés.

- L'âge est lié à l'ancienneté avec un coefficient de 0,24 et avec la nature de contrat avec un coefficient de 0,37. L'âge est forcément lié à l'ancienneté car plus un salarié est âgé plus il a de chance d'avoir de l'ancienneté dans l'entreprise où il travaille. Puis, le lien entre l'âge et la nature du contrat est explicable par le fait qu'un salarié plus âgé a plus de chance d'être en CDI qu'un jeune salarié par exemple. La nature du contrat dépend donc de l'âge.

- Le sexe est corrélé avec le salaire avec un indice de 0,28, avec la rémunération annuelle du salarié (0,26), puis avec le secteur lourd, moyen et léger de l'établissement (0,38).
L'inégalité est encore présente entre les hommes et les femmes, et le genre peut être corrélé avec le salaire. De plus, il n'est pas étonnant de voir une corrélation entre le sexe et le secteur de l'établissement, car la

[23] Le lexique des variables se trouve en **annexe 4.**

tendance à travailler dans un domaine à stress plus ou moins fort est différente selon les hommes ou les femmes.

- La région est liée avec la nature du contrat avec un faible coefficient de 0,14, à la taille de l'établissement (0,15), au statut conventionnel du salarié avec un coefficient de 0,18 et au secteur des entreprises et des établissements avec un indice de 0,19.
Il est logique que certaines régions aient tendance à concentrer les grandes entreprises (Ile de France par exemple), ou encore à concentrer des entreprises à secteur lourd. Aussi, certaines régions peuvent avoir plus de cadres que de non-cadres, en fonction de l'attractivité et de la concentration des grandes entreprises au sein de la région.

- Le salaire est très lié au statut du salarié avec un coefficient de 0,39 et encore plus avec le secteur de l'entreprise (lourd, moyen, léger) avec un indice de 0,61.
Il est évident que le salaire reçu par un salarié est très dépendant de son statut mais aussi du secteur dans lequel il se situe. Un salarié dans un secteur lourd a généralement une charge de travail plus élevée et est donc susceptible d'être mieux rémunéré.

- Le motif de reprise du travail après un arrêt est lié à l'ancienneté (0,17), au motif de fin du contrat (0,19), à la rémunération annuelle du salarié (0,22), à la nature du contrat (0,28) et au statut du salarié (0,28).
La raison pour laquelle le salarié reprend le travail est donc dépendante de son type de contrat et de sa rémunération. Cela s'explique par le fait que la raison de retour au travail dépend beaucoup de la situation de chaque salarié.

- L'ancienneté est liée à la nature du contrat (0,62), au motif de fin de contrat (0,29) et au statut du salarié (0,19) puis au secteur (lourd, moyen, léger) de l'établissement avec un coefficient de 0,23.
Le type de contrat dépend de l'ancienneté du salarié car il peut dépendre de son expérience. Il est logique qu'elle dépende aussi du motif de fin de contrat, car selon l'ancienneté les motivations d'arrêter un contrat ne sont pas les mêmes.

- Le motif de fin de contrat est lié au motif d'arrêt de travail (0,24), à la nature du contrat très fortement (0,87), au statut du salarié (0,29), au secteur de l'établissement (0,44) et à la gestion du contrat (déléguée ou directe) avec un indice de 0,59.
La motivation d'arrêter son contrat peut être liée aux raisons des arrêts (burn-out pour les cadres, inaptitude au poste suite à l'arrêt etc.). Mais aussi, elle est très corrélée à la nature du contrat surtout pour les contrats CDD et intérimaires. Puis, le fait d'être dans un secteur lourd, moyen ou léger influence naturellement la clôture du contrat car le salarié peut être affecté par la pénibilité du secteur dans lequel il se trouve.

- La nature du contrat est liée au type de gestion (directe ou déléguée) avec un indice de 0,44 et avec le secteur (0,25). Nous acceptons que le type de contrat est évidemment corrélé au secteur lourd, moyen ou léger car nous ne trouvons pas les mêmes types de contrats dans les différents secteurs.

- Le statut conventionnel du salarié est lié au motif d'arrêt de travail (0,43). Les cadres et les non-cadres n'ont pas les mêmes motifs pour les arrêts car ils ne vivent pas forcément dans le même confort, surtout dans les branches « aide à domicile » et « propreté » étudiées.

Le nombre de variables dans notre base est important et les corrélations entre les variables sont nombreuses. C'est pourquoi, une analyse en composantes principales est élaborée pour les variables quantitatives et une analyse en composantes multiples est mise en place pour les variables qualitatives. A l'aide de ces analyses, les variables les plus importantes (celles qui contribuent le plus à la variabilité) pour notre modélisation sont ciblées à priori. Puis, pour les variables qui sont très proches en corrélation comme les montants de rémunération, nous pouvons sélectionner une unique variable comme représentante de tout ce groupe de variables. Ainsi, il est possible de résumer en une seule variable les informations qui sont redondantes.

Il est à noter que les corrélations étudiées dans ces analyses se basent sur des relations de linéarité. C'est une étude à priori qui n'est donc pas complète.

2.4.2 Analyse en composantes principales

Voici les résultats de l'ACP, en commençant par ces graphiques.

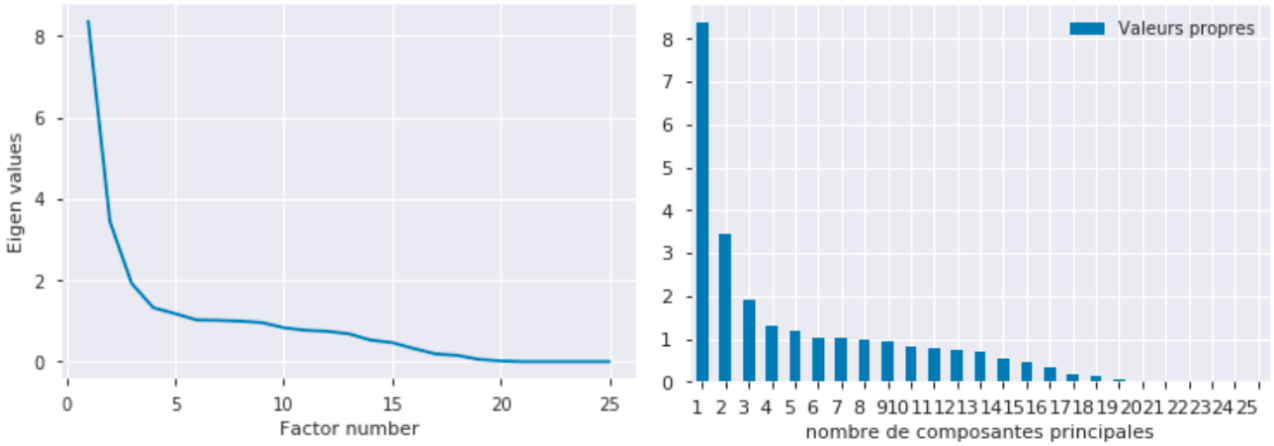


Figure 37 : scree- plot

Ces graphiques appelés « scree plot » tracent la décroissance des valeurs propres en fonction du nombre de composantes principales. Ils permettent de choisir le nombre de facteurs expliquant la variabilité d'après le changement de structure appelé le « coude ». Sur ce scree-plot, nous pouvons observer la présence du « coude » pour un nombre de trois composantes principales. Mais, l'inertie captée par deux composantes principales est déjà très importante puisque que la première composante principale a une valeur propre de 8,36 et la deuxième de 3,44, puis les autres sont inférieures à 2. Nous pouvons donc nous limiter à deux facteurs principaux.

Nous présentons le graphique des corrélations des variables quantitatives avec ces deux axes.

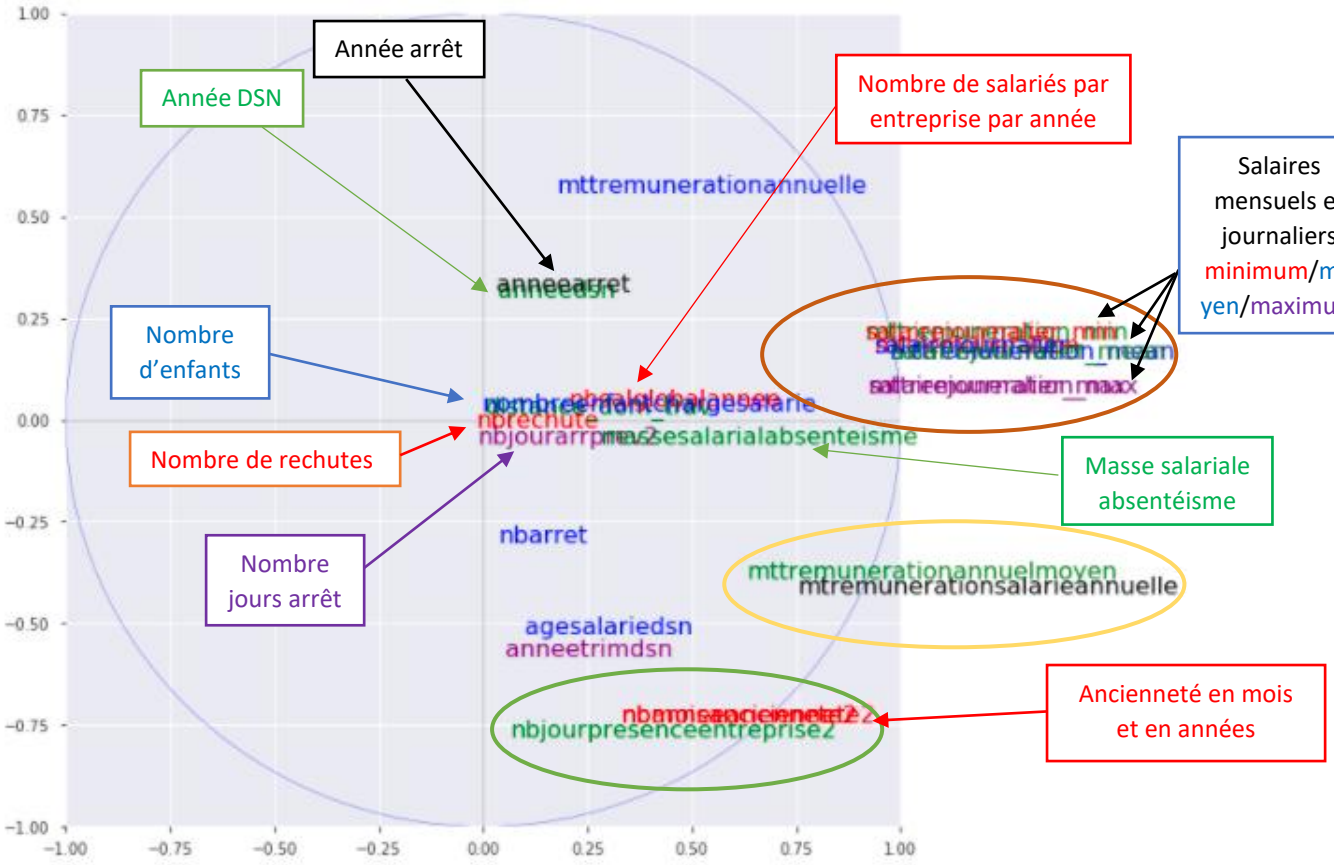


Figure 38 : graphique de l'ACP

Pour lire les corrélations sur ce graphique, il faut savoir que plus une variable est proche de l'axe étudié plus la variable est corrélée à cet axe. Cette corrélation est donnée par sa coordonnée sur l'axe. Puis, pour traduire la corrélation entre les variables, trois choses sont retenues :

- Des variables à proximité sur le graphique sont positivement corrélées ;
- Des variables diamétralement opposées sont corrélées négativement ;
- Des variables formant un angle droit en partant du centre ne sont pas corrélées.

Un point d'attention est à noter : si les variables sont très proches du centre, elles sont souvent mal représentées par le plan factoriel. Nous ne pouvons pas faire confiance à la proximité entre elles et avec les axes pour faire des interprétations.

Nous pouvons déjà remarquer les trois groupes de variables entourés sur ce graphique : le jaune et le rouge représentatifs des revenus, et le vert représentatif du temps en entreprise. Les variables appartenant à un même cercle sont proches et cela montre une forte corrélation entre elles. Parmi chaque groupe de variables entouré, nous pourrions (en théorie) donc en sélectionner seulement une (pour le modèle) qui est représentative des autres corrélées avec celle-ci. Nous nous concentrons uniquement sur les variables de revenu. Pour résumer l'information sur le revenu et sélectionner une variable par groupe (rouge et jaune), nous ciblons les variables les plus contributives aux composantes principales.

Les corrélations des variables avec ces deux composantes principales sont étudiées sur les graphiques en **annexe 5.** Il est possible d'en déduire les contributions à l'aide d'histogrammes pour plus de clarté. La contribution est calculée à partir du carré de la corrélation normé par l'importance de l'axe (la valeur propre).

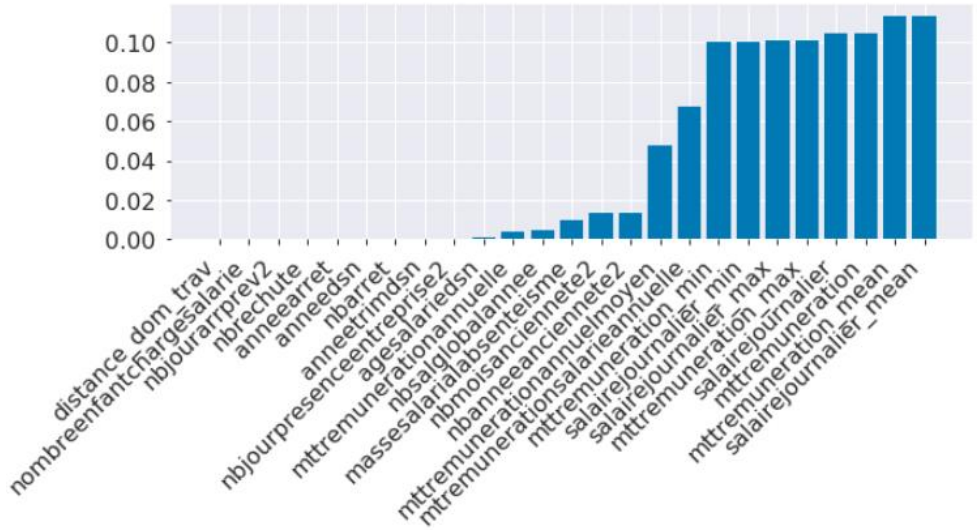


Figure 39 : contributions des variables avec le premier axe

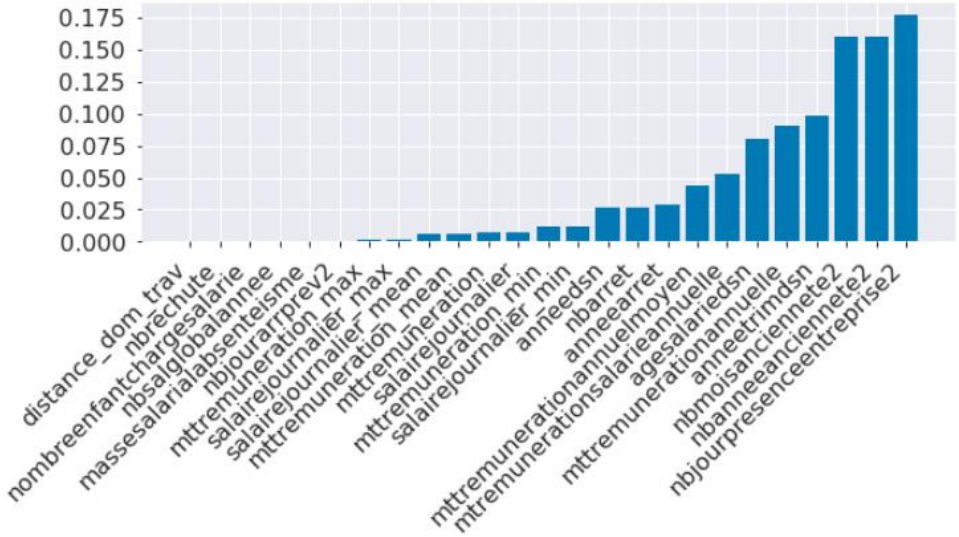


Figure 40 : contributions des variables avec le second axe

Les variables les plus contributives à la première composante principale sont celles qui concernent la rémunération, surtout le salaire journalier moyen par salarié (et la rémunération mensuelle) et les montants de rémunération du salarié pour ses contrats de travail. Le premier axe principal résume donc l'information sur les revenus.

Les variables les plus contributives au second axe principal sont les variables liées à la présence dans l'entreprise comme la durée de présence en entreprise sur une année, l'ancienneté du salarié, les trimestres de déclarations etc.

De là, nous pouvons conclure que les variables les plus importantes pour notre modélisation (celles qui expliquent le plus la variabilité des individus) sont celles liées au revenu avec le revenu moyen par salarié en tête, l'ancienneté et la présence en entreprise, la masse salariale de l'absentéisme et l'âge du salarié.

L'objectif de cette ACP est aussi d'éliminer les variables qui répètent l'information sur le revenu du salarié. Nous pouvons sélectionner une variable pour chaque groupe encerclé par le rouge et par le jaune. En regardant les corrélations, pour le cercle rouge, c'est le montant de rémunération mensuel moyen par salarié qui est le plus corrélé à la première composante principale et qui est donc sélectionné. Puis, pour le cercle

jaune, nous sélectionnons le montant de rémunération annuel du salarié qui est plus corrélé aux composantes principales que le montant de rémunération annuel moyen par entreprise.

2.4.3 Analyse en composantes multiples

Avec l'ACM, les variables qualitatives qui expliquent le plus la variabilité des individus sont obtenues. Cette méthode est moins détaillée que l'ACP et elle ne sert pas à éliminer des variables du jeu de données car dans la matrice de corrélation du V de Cramer, il n'y a pas de groupe de variables (autant imposant que les variables sur le revenu) très corrélées.

Avant de faire l'ACM, les variables sont binarisées. Le nombre très important de modalités nous empêche donc d'interpréter la représentation de celles-ci sur un graphique. Il est cependant possible de regarder les coordonnées (égales aux corrélations) des variables sur les histogrammes en **annexe 7.**

Pour connaître les variables les plus contributives au premier axe et au second axe, nous regardons les variables qui ont les coordonnées les plus importantes.

Les variables les plus liées au premier axe ont un coefficient de corrélation élevé avec le premier axe, c'est-à-dire des coordonnées négatives ou positives élevées : le sexe (masculin), les tranches d'âge (de 16 à 19 ans), la région de l'établissement (Basse Normandie), les salaires (2000€ à 3000€) etc. Puis celles qui ont une corrélation négative sont : le motif d'arrêt de travail (aucun arrêt), le motif de fin de contrat (licenciement) etc.

Les variables les plus contributives au second axe sont : la tranche de rémunération annuelle du salarié (3000€ à 4000€), l'ancienneté (inférieure à 1 an), la gestion déléguée etc. Dans les corrélations négatives, nous observons les variables suivantes : le motif de fin de contrat (licenciement), l'ancienneté (supérieure à 10 ans), les salaires (4000€ à 5000 €) etc.

Nous avons donc effectué une étude à priori des différents liens existants entre les variables et ciblé celles qui contribuent le plus à la variabilité des individus. Il est donc attendu à ce que celles-ci soient discriminantes dans nos modélisations à venir.

Partie 3 : les modèles théoriques

Dans cette partie, la théorie des modèles suivants sera expliquée : CART, Random Forest, XG Boost, Kaplan Meier, GLM et GAM. Puis, les différents outils qui permettent de faire une comparaison seront présentés.

Section 3.1 : l'algorithme CART

3.1.1 Principe

L'algorithme CART (Classification And Regression Trees) a été élaboré en 1984 par Leo Breiman. Il sert à entrainer un arbre décision en partitionnant de manière récursive un échantillon en plusieurs parties. Il peut avoir deux objectifs : la classification ou la régression. Pour une classification, la variable à prédire est binaire ou catégorielle et non ordonnée. L'arbre sert alors à affecter un individu ou un autre objet à une classe. Dans le cas d'une régression, la variable à prédire est quantitative ordonnée et l'arbre est utilisé pour expliquer la variable cible et prédire sa valeur en fonction des variables explicatives.

La notion d'arbre de décision et sa constitution est maintenant défini. Un arbre de décision est composé par définition d'une racine, de branches, de nœuds et de feuilles. Les nœuds sont à la racine des branches et permettent de tester un critère de décision pour ensuite attribuer un nombre de décisions égal au nombre de branches. La racine est le premier critère de division, c'est-à-dire le premier nœud, et les feuilles sont les nœuds terminaux, c'est à dire les classes ou les valeurs finales.

Ensuite, est introduit le concept de l'arbre binaire. Notons que ce type d'arbre est incontournable car il faut savoir que tout arbre peut être traduit en un arbre de décision binaire. Sa particularité se situe au niveau du nœud qui donne naissance à seulement deux branches car le critère de division admet deux réponses : oui ou non.

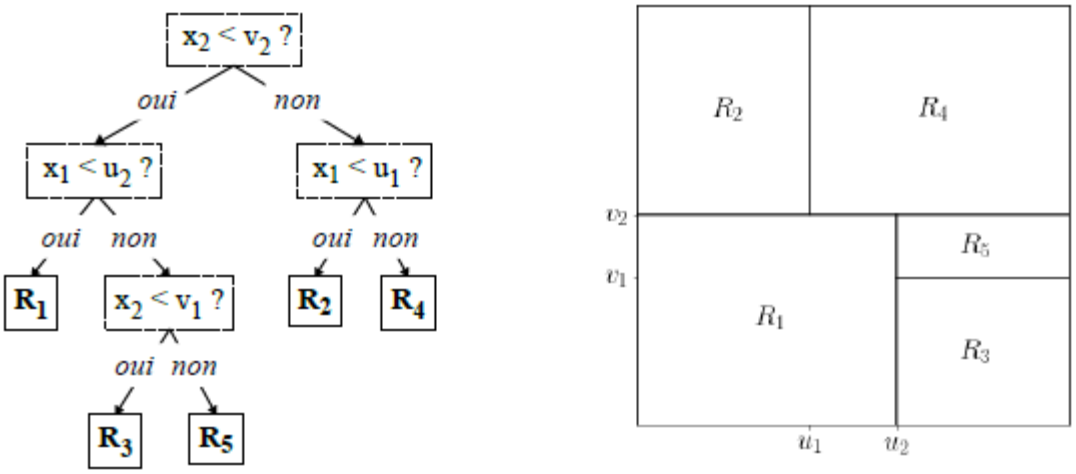


Figure 41 : exemple d'un arbre binaire[25]

A présent, nous pouvons détailler l'algorithme CART qui permet de construire l'arbre de décision optimal. L'algorithme se décompose en plusieurs étapes :

- Définition d'un critère de division ;
- Divisions des nœuds selon des critères réitérés ;
- Mise en place d'un critère d'arrêt de division ;
- Obtention de l'arbre maximal : il contient un nombre de nœuds maximum ;
- Elagage de l'arbre afin d'obtenir un arbre optimal plus flexible et qui ne subit pas les problèmes de surapprentissage.

Les différentes phases de la construction de l'arbre sont détaillées ici.

3.1.2 Critère de division

Le critère de division de l'échantillon n'est pas le même dans le cas où la variable étudiée est catégorielle ou numérique. Si elle est numérique, la division se fait selon une valeur de référence en séparant l'échantillon en fonction de l'infériorité ou la supériorité de la valeur. Si ce n'est pas le cas, la séparation de l'échantillon se fait selon les modalités de la variable, modalité par modalité, en divisant l'échantillon selon la réponse « oui » ou « non » d'égalité à cette modalité.

En termes mathématiques, voici comment se définit ce critère d'arrêt. Soit Y la variable cible que nous voulons prédire, J l'échantillon à chaque nœud, I et K les échantillons obtenus par le critère de division. Les valeurs prédites de la variable réponse sont notées $\hat{y}_I$ et $\hat{y}_K$. L'algorithme recherche la division qui rend le critère de division maximal, c'est-à-dire celle qui minimise l'hétérogénéité des nœuds fils. Une nouvelle itération de l'algorithme est ensuite réalisée pour les deux nœuds fils ainsi construits.

Critère de division : $\sum_{j\in J}(y_j - \hat{y}_J)^2$ - $\sum_{i\in I}(y_i - \hat{y}_I)^2$ + $\sum_{k\in K}(y_k - \hat{y}_K)^2$

[25] Azencott C. (2018) : Introduction au Machine Learning

Avec y_j, y_i et y_k les valeurs réelles observées de la variable cible Y des échantillons J, I et K respectivement.

Ce principe est illustré avec les graphiques suivants montrant le processus de la division optimale d'un échantillon.

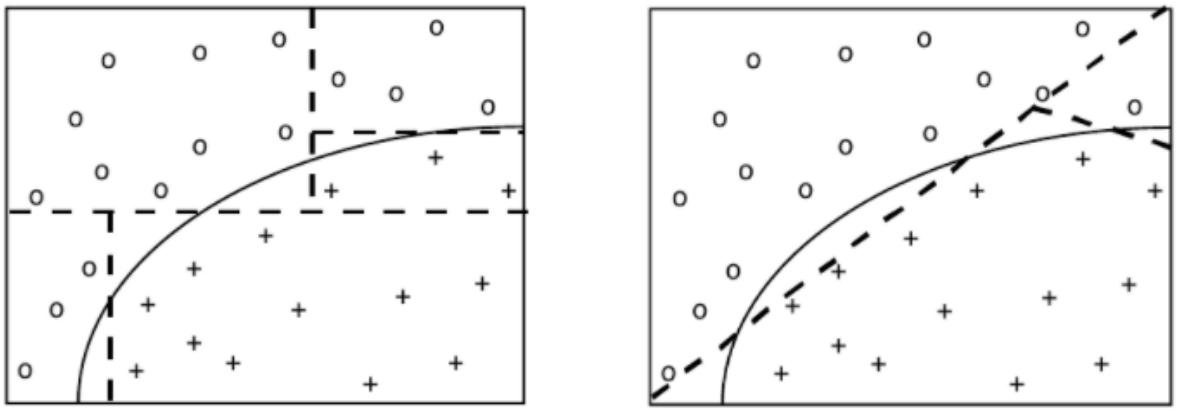


Figure 42 : illustration de la division optimale d'un échantillon[26]

Nous observons ci-dessus à gauche le partitionnement de l'échantillon qui est optimisé par la réitération du critère d'homogénéité maximal dans chaque nœud sur chaque variable. A droite, nous observons le partitionnement cette fois-ci à l'aide de combinaisons linéaires sur certaines variables.

L'étape suivante consiste à limiter les divisions des nœuds au risque d'obtenir un arbre avec trop de divisions.

3.1.3 Critère d'arrêt

Une règle d'arrêt de division des échantillons doit être fixée, sans laquelle nous obtiendrions seulement des feuilles terminales avec des échantillons composés d'un seul objet. La chose à éviter est de faire un partitionnement trop fin qui serait inutile. Avec cette règle d'arrêt, nous pouvons ainsi obtenir un arbre maximal qui décide de s'arrêter à un nombre de nœuds terminaux. La règle peut être par exemple de définir un nœud comme terminal lorsque l'échantillon fils est plus petit qu'un seuil donné ou quand l'augmentation de l'homogénéité d'une division à une autre passe en dessous d'un seuil.

Lorsque l'arbre maximal a été obtenu, la prochaine étape consiste à élaguer l'arbre pour obtenir un arbre optimal.

3.1.4 Technique d'élagage

Cette phase de l'algorithme s'avère primordiale pour éviter le surapprentissage de l'arbre maximal. Il y a deux étapes encore ici. La première consiste à obtenir plusieurs arbres de régression puis de sélectionner celui qui est optimal. Une méthode consiste à utiliser la cross validation sur chaque arbre possible. C'est-à-dire que

[26] Source : http ://cedric.cnam.fr/vertigo/Cours/ml2/coursArbresDecision.html (Apprentissage avec les arbres de décisions)

l'échantillon est séparé en n blocs. n-1 blocs servent comme échantillons d'apprentissage pour entrainer l'arbre, puis, l'autre bloc sert comme échantillon de test pour calculer l'erreur d'estimation. Ceci est répété n fois, à chaque fois en prenant un autre bloc pour l'échantillon de test. En fonction du nombre de feuilles, une erreur relative de prédiction est ainsi obtenue et le nombre de feuilles pour lequel l'erreur est la plus faible est sélectionné. Ceci est illustré sur le graphique ci-dessous. Dans ce cas-ci par exemple, nous nous limitons à 9 feuilles, car c'est là où l'erreur est minimale.

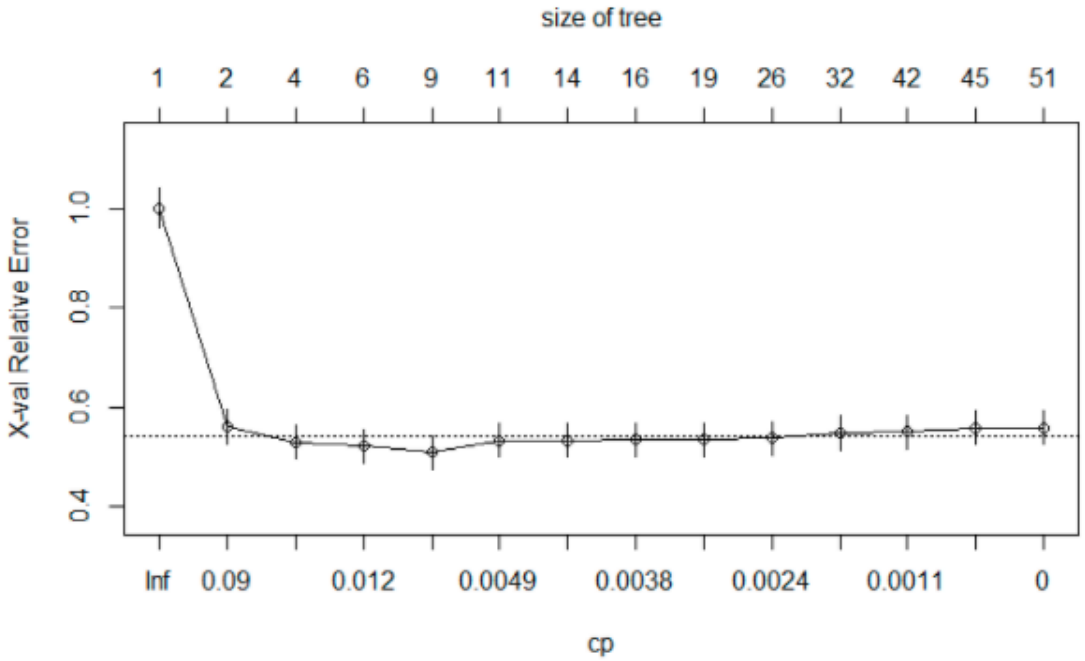


Figure 43 : exemple du tracé de l'erreur de prédiction relative en fonction du nombre de feuilles [27]

3.1.5 Avantages et inconvénients de l'algorithme CART

Quelques avantages de la méthode CART sont listés :

- La méthode est non paramétrique : aucune hypothèse de loi est établie. Donc, les effets non linéaires entre la variable cible et les variables explicatives peuvent être pris en compte ;
- La schématisation de l'arbre rend la compréhension très claire de la classification ;
- L'arbre s'adapte à des valeurs qualitatives et aussi quantitatives.

Mais aussi, il y a quelques inconvénients :

- L'arbre s'adapte trop à la taille de la base d'apprentissage appelée « Train » ;
- Le critère de division dépend trop des variables ayant beaucoup de modalités, ce qui peut biaiser les prédictions, en omettant les variables de faibles modalités ;
- Il y a une certaine complexité dans la phase de l'élagage.

Face aux limites que présente ce modèle de prédiction, nous introduisons l'algorithme des Random Forest qui vient améliorer les problèmes de surapprentissage et de robustesse.

[27] Source : https ://www.lovelyanalytics.com/2016/08/18/un-arbre-de-decision-avec r/?like_comment=566

Section 3.2 : Random Forest

3.2.1 Principe

Les méthodes de forêts aléatoires viennent plus tard que l'algorithme CART et sont élaborées par Brieman dans un but d'amélioration du modèle CART.

Tandis que ce modèle présenté précédemment souffre de problèmes de surapprentissage, les Random Forest permettent de rendre l'arbre plus robuste et plus flexible. Comme nous avons vu dans les faiblesses de la méthode CART, l'arbre dépend trop de la base Train. Il est donc nécessaire de rajouter un aspect aléatoire aux données lors de la construction des arbres. C'est sur cela que se base ce modèle amélioré dont le fonctionnement est expliqué.

La première étape de l'algorithme consiste à faire des échantillonnages aléatoires. Parmi l'échantillon constituant la base d'apprentissage, nous construisons N échantillons aléatoirement de même nombre d'observations et de variables que le Train. Les N arbres de décision sont alors créés à partir de ces N échantillons.

La prochaine étape réside dans le choix du nombre n de variables qui est fixé pour la construction de chaque arbre. Le but est ici de diminuer la variance du modèle choisi à la fin. Les variables choisies pour chaque arbre diffèrent en fonction de l'échantillon. Les variables explicatives ne sont pas jugées de la même importance pour la prédiction selon des jeux de données différents. De ce fait, nous observons des arbres différents qui ne sont pas construits forcément sur les mêmes variables prédictives. Pour chacun d'entre eux, les variables sélectionnées sont celles qui donnent la division optimale des échantillons à chaque nœud.

La dernière étape consiste à obtenir un arbre optimal, celui qui grâce à tous ces arbres combinés obtient un maximum de performance. Ce modèle prédictif optimisé n'est pas obtenu pareillement selon la méthode : classification ou régression. Lors d'une classification, c'est-à-dire pour prédire une variable qualitative, nous retenons comme réponse prédite la classe la plus représentée parmi tous les arbres, nous procédons en réalité à un simple vote. Lors d'une régression, pour prédire une variable quantitative, nous considérons simplement la moyenne des valeurs prédites par chaque arbre.

Pour illustrer le fonctionnement des Random Forest, voici un schéma :

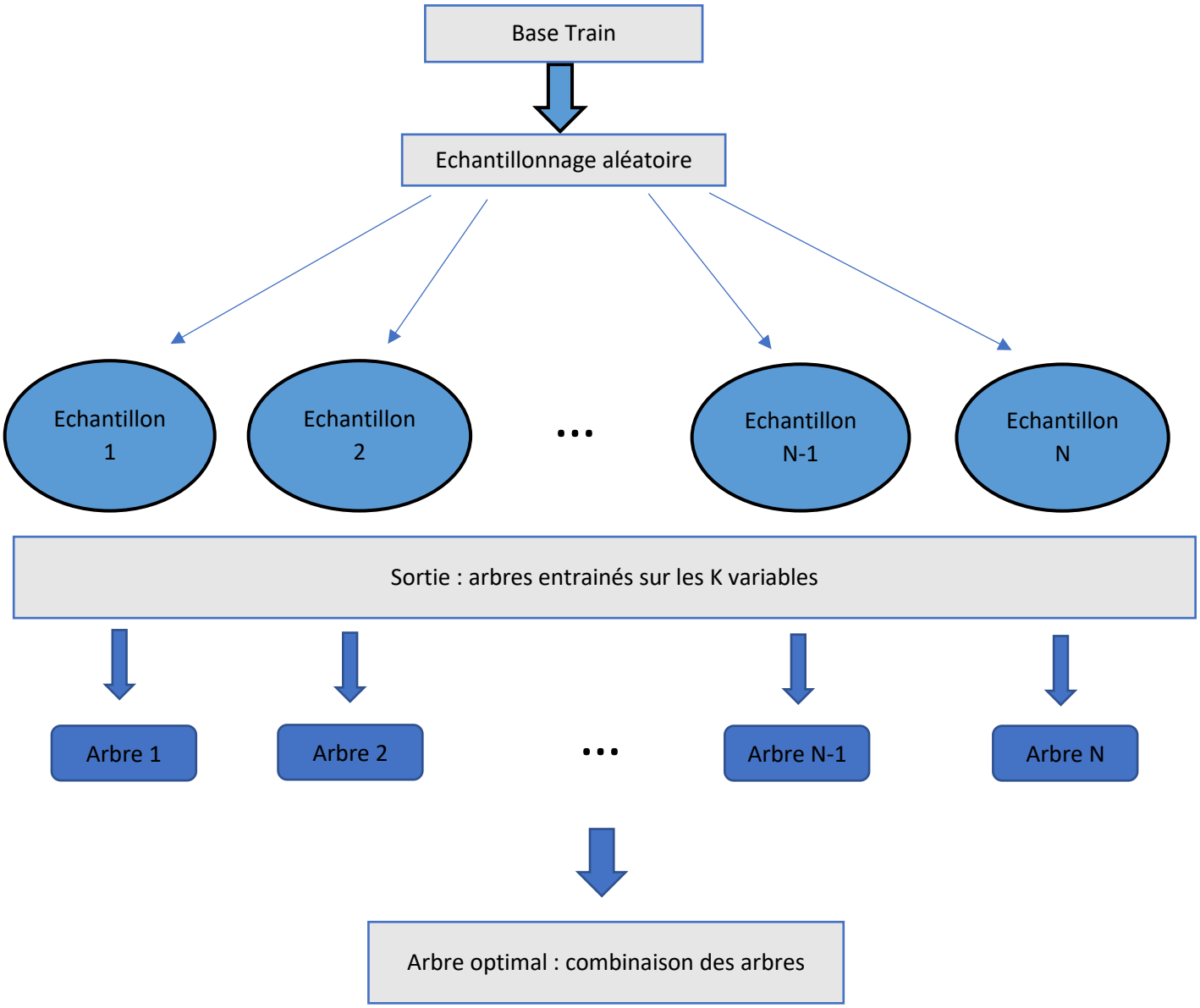


Figure 44 : schéma explicatif du Random Forest

Mais, bien entendu, la prédiction n'est pas tout à fait exacte et il existe une erreur de prédiction : l'Out Of Bag.

3.2.2 L'erreur Out Of Bag

Cette fois ci, l'erreur de prédiction ne se calcule pas à l'aide d'une validation croisée et d'un échantillon test mais se calcule sur toutes les observations qui n'ont pas servi à la constitution des échantillons Bootstrap. L'erreur de prédiction calculée est la suivante :

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

Où la prédiction d'une observation y_i est : $\hat{y}_i = \frac{1}{|J_B|}\sum_{j \in J_B} f(x_i)$

Avec

- m : le nombre d'observations « Out of Bag » ;

- J_B : l'échantillon des valeurs absentes dans l'échantillon Bootstrap ;

- x_i : les variables explicatives ;

- f : les fonctions de prédiction de chaque arbre.

Une fois le modèle élaboré, il est possible de classer les variables par ordre d'importance. Pour mesurer l'importance d'une variable, le processus est le suivant : les modalités de cette variable sont changées de façon aléatoire tout en laissant les modalités des autres variables telles qu'elles sont. Le but est de calculer la conséquence de cette permutation sur les arbres de prédiction. L'importance de la variable est jugée par sa significativité dans les prédictions. Elle-même est calculée selon l'écart d'erreur sur chaque arbre avant et après la modification, puis en faisant la moyenne de cet écart sur les N arbres.

3.2.3 Avantages et inconvénients du Random Forest

Les avantages pour les Random Forest sont listés :

- Amélioration de la flexibilité du modèle, en limitant le surapprentissage de la méthode CART ;
- Précision de prédiction, même si les données sont incomplètes ;
- Baisse de variance par rapport à un arbre de décision seul ;
- Mesure simple de l'importance des variables et de l'erreur de prédiction ;

Ce modèle présente aussi ses limites que voici :

- Nécessité de logiciels puissants ;
- Complexité du modèle ;
- Temps de calcul important ;
- Sensibilité au surapprentissage encore présente.

Il est possible d'améliorer encore le modèle de prédiction et c'est dans cet objectif que la modélisation est approfondie avec une méthode plus performante : le Gradient Boosting Machine.

Section 3.3 : Gradient Boosting

L'algorithme de base de ce modèle, le Boosting, est d'abord présenté.

3.3.1 Le Boosting

Parmi les premiers algorithmes de Boosting, a été inventé par Freund et Schapir en 1996 l'algorithme ADABOOST. Le principe est de combiner les hypothèses faibles pour former une hypothèse forte (week learner et strong learner). Cette méthode est encore une fois basée sur les arbres de décisions et vient y apporter des améliorations du modèle.

Elle consiste en la création d'un arbre simple, c'est-à-dire avec une seule division et deux feuilles. Un poids est attribué pour chaque observation mal prédite. A partir de cette observation pondérée est alors construit un autre arbre et cette étape est répétée pour chaque nouvel arbre créé. Ce qui nous donne une séquence d'arbres. Voici ci-dessous une illustration du processus.

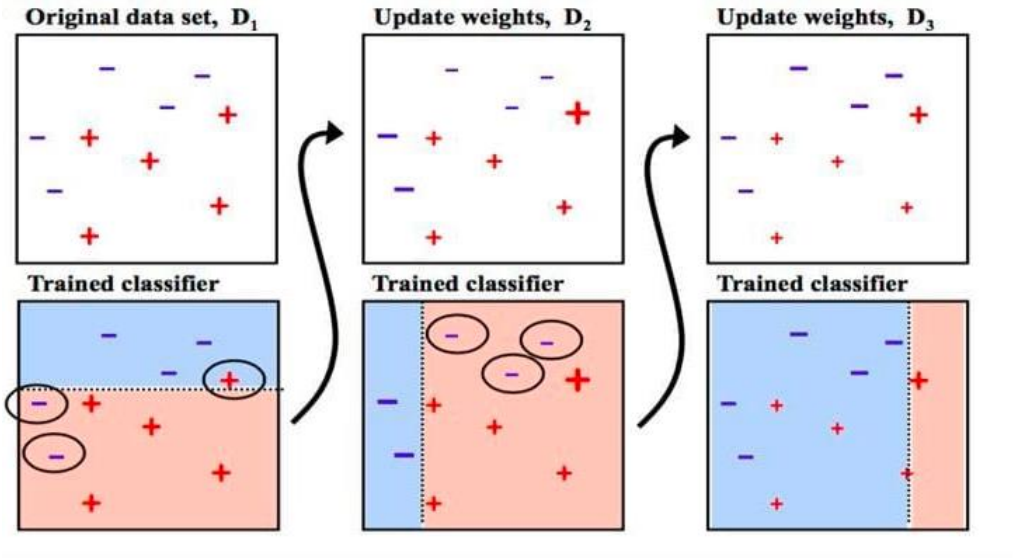


Figure 45 : schéma de la suite des classifications selon les observations pondérées (Boosting)[28]

Comme le montre le schéma ci-dessus, le même poids est attribué à chaque observation lors de la première modélisation D1. Puis, les poids des observations mal classées augmentent en fonction du niveau d'erreur de classification et le prochain modèle D2 se base sur les données pondérées par ces erreurs. Ainsi, nous observons qu'à chaque étape de l'algorithme, l'échantillon se divise en deux classes en prenant en compte les erreurs de prédiction.

La force de cet algorithme se situe là : grâce à tous ces classifieurs faibles, un classifieur fort est obtenu par pondération de ceux-ci. Se pose enfin la question de l'agrégation des classifieurs faibles entre eux. Ils sont combinés de la manière suivante :

$$H_{final}\ (x) = \text{sign}\left(\sum_{t=1}^{T}\ \alpha_t\ h_t(x)\right)$$

Avec :

-H_{final} : classifieur fort ;

[28] Source : https ://iq.opengenus.org/gradient-boosting/

- α_t : pourcentage d'observations biens classées pour un classifieur faible n° t ;

- h_t : classifieur faible ;

- T : nombre de classifieurs faibles ;

- sign : signe du poids du classifieur obtenu à la fin.

Comme pour le Random Forest, nous procèdons à un vote pour chaque observation qui est multiplié par le poids d'un classifieur (qui représente le pourcentage d'observations biens classées).

Le Gradient Boosting Machine part du même principe qu'ADABOOST, sauf que lors des votes, tous les classifieurs faibles ont le même poids.

3.3.2 Le Gradient Boosting Machine

Cette partie est inspirée du mémoire de Didrik NIELSEN.[29]

L'algorithme du Gradient Boosting Machine permet tout comme ADABOOST de faire des classifications et des régressions. Mais, dans cette méthode, le but est de minimiser la fonction Loss qui représente l'écart entre la variable observée y et sa valeur prédite $f(X)$ en fonction des variables d'apprentissage $X = (x_1 , x_2 , \dots , x_P)$.

La fonction Loss peut s'écrire de plusieurs manières :

- $L\big(y, f(X)\big) = (y - f(X))^2$
- $L\big(y, f(X)\big) = | y - f(X) |$
- $L\big(y, f(X)\big) = \mathbb{1}_{y \neq f(X)}$

L'objectif est de minimiser l'erreur d'estimation, nous cherchons donc $\tilde{f}$ telle que :

$$\tilde{f} = \arg \min_f \mathbb{E} \left[L\big(y, f(X)\big) \right]$$

Grâce au GBM, une estimation de cette fonction $\tilde{f}$ est obtenue en pondérant les classifieurs faibles :

$$\tilde{f}(X) = \sum_{i=1}^{M} \gamma_i \, h_i(X) + C$$

Avec :

- C : une constante ;
- h_i : les classifieurs faibles ;
- M : le nombre de classifieurs faibles ;
- γ_i : les poids des classifieurs faibles.

L'algorithme part de la première valeur de la fonction f_0 qui minimise la Loss pour le premier classifieur faible, c'est-à-dire sur la moyenne des écarts sur toutes les observations i ∈ {1, …, n} :

[29] Didrik NIELSEN (2016) : Tree Boosting With XGBoost : Why Does XGBoost Win "Every" Machine Learning Competition ?

$$f_0(X) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

Ensuite, à l'étape d'après il recommence la même chose en ajoutant cette fois-ci la fonction estimée à l'étape d'avant, et ceci pour chaque itération $m \in \{1,\dots, M\}$:

$$f_m(X) = f_{m-1}(X) + \arg\min_{f} \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i)) + g(x_i)$$

avec g une fonction qui est dérivable et monotone.

Nous notons que ce problème d'estimation est assez complexe notamment à cause de la fonction g et le GBM qui se base sur des pseudo-résidus calculés à chaque étape d'itération (ce qui est appelé une descente de gradient). C'est pourquoi, le GBM suit un algorithme bien précis afin de résoudre ce problème.

Voici les différentes étapes du GBM pour résoudre le problème d'estimation :

- **Etape 1** : Initialisation

  $f_0(X) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$

- **Etape 2** :

Pour chaque itération $m \in \{1,\dots, M\}$:

- Calcul des résidus

  $r_{i,m} = -\left[\frac{d\, L(y_i, f(x_i))}{df(x_i)}\right]_{f(x)=f_{m-1}(x)}$

- Calcul d'un classifieur h_m sur un échantillon $(x_i, r_{i,m})_{i=1}^{n}$
- Calcul de γ_m tel que

  $\gamma_m = \arg\min_{f} \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \gamma\, h_m(x_i))$

- Agrégation des fonctions f tel que : $f_m(X) = f_{m-1}(X) + \gamma_m\, h_m(X)$

- **Etape 3** : Sortie
  Estimation de $\tilde{f} = f_M(X)$

Pour aller plus loin et obtenir un modèle plus performant, le XGBoost est présenté.

3.3.3 Le XGBoost

Cet algorithme est connu pour ses fortes performances notamment à travers les concours Kaggle. Ce modèle est une version du GBM mais apporte des améliorations dans la performance et dans le temps d'exécution du programme grâce à la parallélisation des traitements. Il a aussi d'autres avantages par rapport au GBM :

- Il permet d'implémenter des modèles linéaires qui complètent la méthode par le biais de classifieurs faibles, alors que le GBM utilise seulement les arbres de régression. Cela permet donc de diminuer le biais en gardant une variance raisonnable.
- Tandis que le GBM utilise la fonction Loss pour évaluer la qualité de prédiction du modèle, le XGB introduit en plus une fonction de régularisation qui permet de contrôler la robustesse et qui pénalise la complexité du modèle.
- Le XGB prend en compte les valeurs manquantes de la base de données et elles n'affectent donc pas le gradient.
- Alors que l'algorithme du GBM s'arrête dès qu'une dégradation du modèle est constatée à cause d'une division, le XGB réalise toutes les divisions puis procède à un élagage. Ainsi, il capte les effets négatifs qui sont suivis d'effets positifs.

Voici un schéma qui illustre comment fonctionne le XGB :

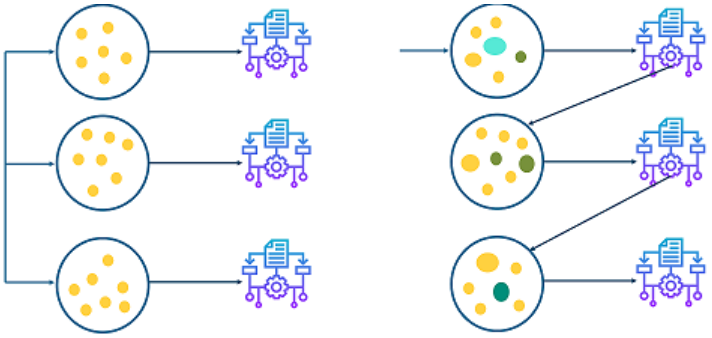


Figure 46 : schéma illustrant le fonctionnement du XGB[30]

Ce schéma permet de comprendre la parallélisation des tâches par l'algorithme du XGB tout en effectuant le Boosting expliqué précédemment.

Le but va être à chaque étape de l'algorithme de minimiser la fonction objective à l'aide d'un arbre supplémentaire. Cette fonction est la somme de la fonction Loss (expliquée dans le paragraphe 3.3.2) et de la fonction de régularisation.

Explication dans le détail des calculs

Soit M_k l'arbre construit à la $k^{ème}$ itération, la prédiction à l'étape k est : $f_m(x_i) = f_{m-1}(x_i) + \eta\, \delta_m(x_i)$

où $\delta_m = -\sum_{i=1}^{n} r_{i,m}$ et η est la paramètre qui contrôle le surapprentissage.

L'objectif de la fonction de régularisation étant de limiter le surapprentissage, il est ajouté à la fonction Loss un terme de pénalité pour la complexité du modèle :

$$\Omega(\delta) = \alpha|\delta| + \frac{1}{2}\lambda||W||^2$$

Avec $|\delta|$ le nombre de feuilles de l'arbre δ et W le vecteur des valeurs des feuilles.

En fait, ce terme peut être vu comme la combinaison d'une régularisation Ridge $\boldsymbol{\lambda}$ et d'une pénalité Lasso α.

Nous notons Obj la fonction objective et elle est exprimée selon la formule suivante :

[30] Source : https ://www.i2tutorials.com/machine-learning-tutorial/bagging-boosting/

$$Obj(f) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega\,(\delta_m)$$

Maintenant, notons $f_{m-1}(x_i)$ l'estimateur de la i -ème observation à l'étape m. Cela revient donc à minimiser :

$$\gamma_m = \sum_{i=1}^{n} L\big(y_i, f_{m-1}(x_i) + \delta_m(x_i)\big) + \Omega\,(\delta_m)$$

Ensuite, γ_m peut être écrit sous la forme d'un développement de Taylor au second degré, ce qui donne :

$$\gamma_m \approx \sum_{i=1}^{n} \left[L\big(y_i, f_{m-1}(x_i)\big) + g_i\delta_m(x_i) + \frac{1}{2}h_i{\delta_m}^2(x_i)\right] + \Omega\,(\delta_m)$$

Avec $g_i = \left[\frac{d\,L(y_i, f_{m-1}(x_i))}{df_{m-1}(x_i)}\right]$ et $h_i = \left[\frac{d^2\,L(y_i, f_{m-1}(x_i))}{df_{m-1}(x_i)}\right]$ les dérivées première et seconde de la fonction Loss.

Si les constantes sont supprimées, nous obtenons finalement la fonction objective suivante :

$$\gamma_m \approx \sum_{i=1}^{n} \left[g_i\delta_m(x_i) + \frac{1}{2}h_i{\delta_m}^2(x_i)\right] + \Omega\,(\delta_m)$$

Notons $I_j = \{i |\ q(x_i) = j\}$ comme la définition de l'appartenance de l'observation x_i à la feuille j. La fonction devient donc :

$$\gamma_m \approx \sum_{i=1}^{n} \left[g_i w_q(x_i) + \frac{1}{2}h_i {w_q}^2(x_i)\right] + \alpha|\delta| + \frac{1}{2}\lambda \sum_{j=1}^{|\delta|} {w_j}^2$$

$$\approx \sum_{j=1}^{|\delta|} \left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) {w_j}^2\right] + \alpha|\delta|$$

Or cette forme quadratique vérifie :

$$argmin_x\ G_x + \frac{1}{2}\,H_x^2 = -\frac{G}{H},\ H > 0 \ \text{ et } \ \min_x G_x + \frac{1}{2}\,H_x^2 = -\,\frac{1}{2}\frac{G^2}{H}$$

Avec $G_j = \sum_{i\in I_j} g_i$ et $H_j = \sum_{i\in I_j} h_i$

Pour un arbre de constitution q(x), nous avons $w_j^* = \frac{G_j}{H_j + \lambda}$ le poids optimal associé à la feuille j.

Ainsi, il est possible d'en déduire la valeur de la fonction objective :

$$Obj = -\frac{1}{2}\sum_{j=1}^{|\delta|} \frac{G_j^2}{H_j + \lambda} + \alpha|\delta|$$

Pour déterminer la meilleure division des nœuds sur un arbre avec uniquement des divisions en deux feuilles, l'algorithme calcul le gain de la fonction objective de manière linéaire pour chaque nœud avec l'expression suivante :

$$Gain = -\frac{1}{2}\left[\frac{G_G^2}{H_G + \lambda} + \frac{G_D^2}{H_D + \lambda} - \frac{(G_G + G_D)^2}{H_G + H_D + \lambda}\right] + \alpha$$

Avec

- $\frac{G_G^2}{H_G + \lambda}$ le poids associé à la branche gauche ;
- $\frac{G_D^2}{H_D + \lambda}$ le poids associé à la branche droite ;

- $\frac{(G_G + G_D)^2}{H_G + H_D + \lambda}$ le poids associé au nœud avant division ;
- α le coût de complexité provoqué par l'ajout d'une feuille.

Lorsque que nous arrivons à un arbre optimal, l'algorithme fait un élagage des nœuds qui ont un gain négatif en allant du nœud le plus profond au moins profond.

3.3.4 Avantages et inconvénients du XGB

Les principales forces de cette méthode sont listées :

- Les arbres sont construits en prenant en compte l'erreur d'estimation, la fonction Loss ;
- Le modèle peut s'appuyer sur différentes fonctions de perte ;
- Il présente une bonne robustesse ;
- Les effets non linéaires peuvent être captés.

Toutefois, la méthode présente certaines limites :

- Nous observons la présence de beaucoup de paramètres qui augmentent la complexité du modèle ;
- Les algorithmes nécessitent des ordinateurs puissants pour les calculs ;
- Une sensibilité au surapprentissage est encore présente ;
- L'interprétation de ce modèle n'est pas évidente.

Section 3.4 : Kaplan Meier

3.4.1 La méthode de Kaplan Meier

La méthode de Kaplan Meier a été conçue par Edward Kaplan et Paul Meier en 1958. Elle permet d'estimer de manière non paramétrique une fonction de survie en passant par l'estimation de taux bruts. Et plus particulièrement, elle vient pallier le problème de censures et de troncatures. Lors d'une modélisation sur une population étudiée pour une période d'observation fixée, certains individus peuvent être perdus en cours de route. Nous pouvons aussi observer un individu seulement après la date de début d'observation. Afin de bien comprendre ces notions, ces termes de censure et de troncature sont définis.

Dans ces définitions, nous nous intéressons à un événement touchant un individu.

Censure

- Censure à droite : c'est quand l'individu n'a pas été touché par l'événement pendant la période d'observation car il a quitté la population avant la fin. Exemple : décès pendant la période d'observation.

- Censure à gauche : nous savons que l'individu a été touché par l'événement qui nous intéresse mais la date de l'événement n'est pas connue.

Troncature

Nous parlons ici de la non prise en compte d'individus dans l'étude, les données sont alors tronquées.

- Troncature à droite : c'est quand une sélection de la population est faite sur la période d'observation. Nous prenons en compte uniquement les individus respectant une condition évènementielle.
- Troncature à gauche : nommée aussi « entrée tardive », c'est dans la situation où les individus ont subi l'événement avant la période d'observation, alors ils ne sont pas pris en compte.

Kaplan Meier permet alors d'estimer la fonction de survie empirique en intégrant les données censurées à droite et tronquées à gauche. Le principe est de séparer la période d'observation en plusieurs échantillons homogènes. Le partitionnement se fait tranche d'âge par tranche d'âge : [âge x ; âge x+1]. A l'intérieur de ces intervalles, le découpage est affiné en créant des sous parties délimitées par des évènements : une censure, une troncature ou un décès. Nous créons ainsi un intervalle sur lequel nous nous basons pour estimer la fonction de survie, où les dates $X_{(i)}$ sont triées dans l'ordre croissant.

La fonction de survie empirique est alors représentée par une fonction escalier. Elle est d'abord égale à 1, puis descend d'un cran à chaque fois. La descente correspondant au rapport du nombre d'individus sortis à la date $X_{(i)}$ sur le nombre de personnes présentes jusqu'à cette date. Les sauts se font aux sorties non censurées.

Pour la suite, notons pour $i \in [1\ ;\ n\ ]$:

- T_i : la durée de survie, c'est-à-dire la durée allant jusqu'à l'événement de l'individu ;
- C_i : la durée de survie jusqu'à la censure ;
- X_i : inf(T_i, C_i) ;
  $X^*_{(i)}$: les valeurs de X_i distinctes ordonnées ;
- δ_i : l'indicatrice qui indique la présence de censure ou non : $\delta_i = 1_{T_i \leq C_i}$. ;
- d_i : le nombre d'individus sortis en $X^*_{(i)}$: $d_i = \sum_{j=1}^{n} 1_{X_j = X^*_{(i)}}\ \delta_i$;
- r_i : le nombre d'individus susceptibles de sortir en $X^*_{(i)}$: $r_i = \sum_{j=1}^{n} 1_{X_j \geq X^*_{(i)}}$.

Nous avons alors l'estimateur de Kaplan Meier suivant :

$$\hat{S}_{KM}(t) = \prod_{i=1:X^*_{(i)} \leq t}^{n} \left(1 - \frac{d_i}{r_i}\right) \text{ (1)}$$

Mais il peut aussi s'écrire, pour t >0 :

$$\hat{S}_{KM}(t) = \prod_{i:X_{(i)} \leq t} \left(1 - \frac{\delta_i}{n-i+1}\right) \text{ (2)}$$

Où les observations $(X_{(i)})_{i\in[1;\ n]}$ sont ordonnées selon l'ordre lexicographique sur $(X_i, 1-\delta_i)$, avec $\delta_i = 1_{T_i \leq C_i}$. Notons que lorsqu'il y a égalité entre les valeurs X_i, la valeur non censurée est placée en premier. Ainsi, nous obtenons $X_{(1)} \leq \ldots \leq \ldots X_{(n)}$ avec les différentes valeurs $1-\delta_{(1)} \leq \ldots \leq 1-\delta_{(n)}$ associées à cet ordre.

Cela est illustré par un exemple de construction de l'estimateur.

Tableau des observations brutes

Individus	Troncatures à gauche τ_i	X_i	Censures δ_i
1	2	7	1
2	2	5	0
3	1	2	1
4	3	4	1
5	1	3	1
6	1	6	0
7	1	2	0
8	3	6	1
9	4	10	0
10	1	2	0
11	1	4	0
12	1	2	1
13	2	8	0
14	4	5	0
15	1	2	0
16	2	8	1
17	1	3	1
18	2	5	1
19	1	2	0
20	1	4	0
21	1	2	1

Pour utiliser la formule de Kaplan Meier (2), les durées sont triées dans l'ordre lexicographique $(X_i, 1-\delta_i)$. Nous obtenons alors pour les 4 premières valeurs :

Numéros d'observations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**…**
$\boldsymbol{min(\tau_i)}$	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	
$\boldsymbol{X_{(i)}}$	2	2	2	2	2	2	2	3	3	4	4	4	5	5	5	…
1-$\boldsymbol{\delta_{(i)}}$	0	0	0	0	1	1	1	1	1	0	0	1	0	0	1	…

Ensuite, la formule de Kaplan Meier peut être mise en application. Mais attention, nous nous intéressons dans ce cas-ci à modéliser la survie sachant la troncature des durées, c'est-à-dire nous voulons estimer la fonction de survie suivante : $S(t|min(\tau_i)) = P(T \geq t | \tau \geq min(\tau_i)) = \frac{S(t)}{S(min\,(\tau_i))}$ pour $t \geq min(\tau_i)$. Le but est alors d'obtenir l'estimation : $\frac{\hat{S}_{KM}(t)}{\hat{S}_{KM}(min(\tau_i))}$

Les résultats sont les suivants :

Pout $t < 2$: $\hat{S}_{KM}(t) = 1$;

Pour $t \in [2\,;3[$: $\hat{S}_{KM}(t) = (1 - \frac{1}{21-1+1}) \times (1 - \frac{1}{20-1+1}) \times (1 - \frac{1}{19-1+1}) \times (1 - \frac{1}{18-1+1}) / \hat{S}_{KM}(1) = 0{,}81$;

Pour $t \in [3\,;4[$: $\hat{S}_{KM}(t) = 0{,}81 / \hat{S}_{KM}(1) = 0{,}81$ car toutes les données sont censurées ;

Pour $t \in [4\,;5[$: $\hat{S}_{KM}(t) = 0{,}81 \times (1 - \frac{1}{21-10+1}) \times (1 - \frac{1}{21-11+1}) / \hat{S}_{KM}(1) = 0{,}68$;

Pour $t \in [5 ; 6[: \hat{S}_{KM}(t) = 0{,}68 \times (1 - \frac{1}{21-13+1}) \times (1 - \frac{1}{21-14+1}) / \hat{S}_{KM}(2) = 0{,}53 / 0{,}81 = 0{,}65$;

...

Pour $t \in [10 ; +\infty[: \hat{S}_{KM}(t) = 0$.

Finalement, nous obtenons le graphique en escalier suivant qui représente l'estimation de la fonction de survie partielle :

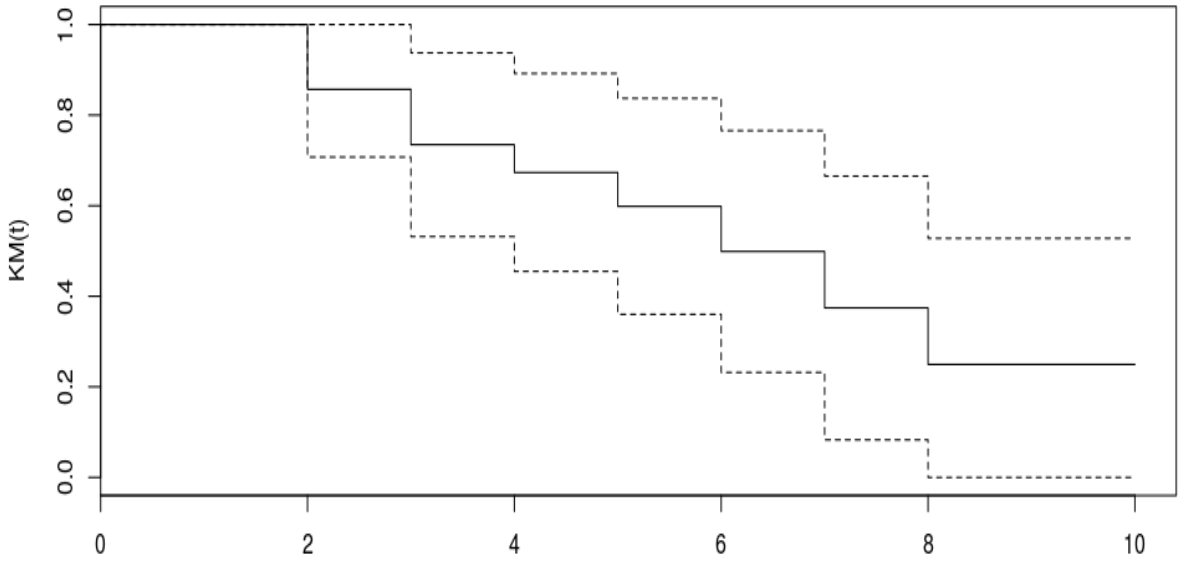


Figure 47 : extrait de l'estimation de la fonction de survie par Kaplan Meier

3.4.2 Avantages et inconvénients de la méthode de Kaplan Meier

Les avantages de cette méthode sont les suivants.

- Le modèle prend en compte les données tronquées et censurées ;
- Le modèle est non paramétrique donc il n'y a aucune hypothèse sur les lois.

L'inconvénient de cette méthode est qu'il faut connaître toutes les dates d'entrées et de sorties du portefeuille.

Section 3.5 : modèles additifs et linéaires généralisés

3.5.1 Modèle GLM

La régression GLM permet d'expliquer une variable réponse par plusieurs variables explicatives. Le GLM se définit en fait par la caractérisation de ces trois éléments :

- La composante principale : la variable ou le vecteur Y à expliquer (dans la suite, par simplicité, nous prendrons Y comme valeur et non comme vecteur). Elle suit une loi qui appartient à la famille exponentielle ;
- La composante systématique η : c'est le prédicteur linéaire. Pour un vecteur de variables explicatives $X = (X_1, \dots, X_p)$et un vecteur à estimer $\beta = (\beta_1, \dots \beta_P)$, voici le prédicteur : $\eta(X) = \sum_{i=1}^{P} X_i * \beta_i$
- La fonction de lien g qui lie la composante systématique et la composante principale comme ceci : $\eta(X) = g(\mu)$ où $\mu = \mathbb{E}[Y]$. Elle doit être monotone et dérivable et correspond souvent à la fonction identité, inverse ou log.

Le GLM repose donc sur trois hypothèses principales :

- Dans le cas où la variable réponse est un vecteur $Y = (Y_1, \dots, \dots, Y_n)$, les Y_i sont des variables indépendantes ;
- La relation de base du modèle est la suivante : $\eta = X \times \beta$;
- La fonction de lien g donne l'expression suivante, avec la possibilité de rajouter un offset ξ : $g(\mu) = \eta + \xi$.

Afin d'utiliser ce modèle, il faut émettre plusieurs hypothèses :

- La loi que la variable Y suit ;
- Les variables explicatives à intégrer dans le modèle ;
- Le type de la fonction lien qui lie les variables explicatives à la variable cible.

La variable réponse Y doit suivre une loi appartenant à la famille exponentielle. C'est à dire une loi de densité qui dépend des paramètres θ et ϕ comme ceci :

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

Avec les éléments suivants :

- a (.) une fonction de $\mathbb{R}$->$\mathbb{R}^*$, b (.) une fonction de $\mathbb{R}$->$\mathbb{R}$ de classe C^2 inversible et c (.) une fonction de $\mathbb{R}^2$-> $\mathbb{R}$;
- θ un paramètre réel naturel ;
- ϕ un paramètre réel de dispersion ;
- Y la variable représentant la cible.

Il s'ensuit les propriétés suivantes qui découlent de l'appartenance à cette famille. Si Y suit une loi exponentielle, nous avons alors :

$$\mathbb{E}[Y] = b'(\theta) = \mu \quad \text{et } \mathbb{V}[Y] = b''(\theta) \times \phi.$$

Notons qu'il est possible d'ajouter un poids ω à la variable à expliquer. Nous obtenons alors la densité suivante :

$$f(y;\omega|\,\theta,\phi) = exp\left[\frac{y*\theta - b(\theta)}{\frac{a(\phi)}{\omega}} + c(y,\phi)\right]$$

Pour illustrer cette famille, il est donné quelques exemples de lois appartenant à la famille exponentielle et leurs caractéristiques.

Tableau de lois appartenant à la famille exponentielle et les paramètres caractéristiques[31]

Lois	Notation	θ	$b(\theta)$	ϕ	ω	$a(\phi)$
Binomiale négative	$\frac{B(n,\pi)}{n}$	$\theta = ln\left(\frac{\pi}{1-\pi}\right)$	$b(\theta) = ln(1+e^{\theta})$	ϕ=1	ω=n	$a(\phi)=\frac{1}{n}$
Poisson	$P(\lambda)$	$\theta = ln(\lambda)$	$b(\theta) = e^{\theta}$	ϕ=1	ω=1	$a(\phi)$=1
Exponentielle	$Exp(\lambda)$	$\theta = \frac{1}{\lambda}$	$b(\theta) = ln(\theta)$	ϕ=1	ω=1	$a(\phi)$=-1
Normale	$N(\mu,\sigma^2)$	$\theta = \mu$	$b(\theta) = \frac{\theta^2}{2}$	$\phi=\sigma^2$	ω=1	$a(\phi)=\sigma^2$
Gamma	$G(a,\lambda)$	$\theta = \frac{1}{a\lambda}$	$b(\theta) = ln(\theta)$	$\phi = -\frac{1}{a}$	ω=1	$a(\phi) = -\frac{1}{a}$

Chaque loi est spécifique et est choisie en fonction de la variable à expliquer du modèle.

- La loi Normale est la loi de base pour effectuer un modèle linéaire classique. Il est représenté par cette équation : $Y = X \times \beta + \varepsilon$, avec Y la variable aléatoire cible à modéliser, X le vecteur des variables explicatives, β le vecteur colonne des paramètres à estimer et ε le vecteur colonne des erreurs. Aussi, le modèle classique repose sur deux hypothèses :
  - Hypothèse sur la linéarité : Y et X sont liés par une corrélation linéaire ;
  - Hypothèses sur l'erreur : les erreurs sont indépendantes, identiquement distribuées et suivent une loi normale centrée.
- La loi de Poisson convient parfaitement pour modéliser une fréquence ou un processus de comptage.
- La loi Gamma est adoptée afin de modéliser des valeurs strictement positives.
- La loi Binomiale est utilisée pour expliquer des variables binaires.

Une fois la loi choisie, viennent les étapes d'estimation des paramètres dans un ordre à suivre. Tout d'abord, ϕ le paramètre de dispersion est estimé. Puis, le vecteur $\beta = (\beta_1, \dots \beta_P)$ est estimé grâce à la méthode du maximum de vraisemblance. Grâce à cela, nous obtenons $\eta(X) = g(\mu)$ et $\mu = g^{-1}(\eta(X))$ l'espérance attendu de Y (la valeur prédite). Lorsque la valeur de μ est connue, nous pouvons calculer $\theta = (b')^{-1}(\mu)$, pour aboutir enfin à la variance de Y : $\mathbb{V}[Y] = b''(\theta) \times \phi$.

Après les GLM, nous expliquons les modèles GAM qui sont basés sur les mêmes principes que les GLM mais avec plus de complexité.

3.5.2 Modèle GAM

Les modèles GAM ont été conçus par T. Hastie et R. Tibshiranie en 1990. Ceux- ci ont été créés afin d'apporter des améliorations au modèle GLM. L'innovation réside dans l'intégration de l'additivité au sein de la composante. La méthode est définie par cette relation :

$$g(\mathbb{E}[Y]) = \mu = \alpha + \sum_{j=1}^{p} f_j(X_j)$$

[31] BABACAR SECK (2006) Estimation pour les modèles linéaires généralisées : Approche marginale, approche conditionnelle et application

Où X_j (j allant de 1 à p) sont les variables descriptives, f_j sont des fonctions prenant en entrée ces variables (un nombre de variables allant de 1 à p) et g est la fonction de lien.

L'objectif est ici d'estimer les fonctions f. Plusieurs méthodes d'estimation sont alors possibles : par des splines cubiques de lissage, par des fonctions de régression locales pondérées, par des fonctions polynomiales, par des trigonométriques etc.

La fonction f vérifie les relations suivantes :

- $Y = f(X) + \varepsilon$ où (ε_t) est un processus indépendant qui suit une loi normale centrée ;
- $\mathbb{E}[Y \mid X = x] = f(X)$.

3.5.3 Avantages et inconvénients du modèle GAM

Les forces de cet algorithme sont les suivantes :

- Les effets non linéaires des variables explicatives et de la variable cible sont captés ;
- Le modèle permet de rajouter des régularisations ou des pénalisations.

Cependant, le modèle présente certaines limites :

- Le lien entre les variables doit être défini ;
- Il y a des problèmes de convergence.

Après avoir décrit tous ces modèles, leurs différents critères sont comparés.

Section 3.6 : Comparaison des modèles

Nous pouvons comparer les modèles GAM, XGBoost, CART et RF utilisés pour les modélisations qui ne fonctionnent pas de la même façon. La méthode de Kaplan Meier qui ne prend en compte qu'une variable pour le calcul des taux est l'approche classique utilisée par les actuaires pour modéliser les arrêts de travail par exemple. Pour prendre en compte les différentes variables/ caractéristiques d'un salarié, nous pouvons adopté le GAM qui est plus facilement interprétable et habituellement utilisé. Puis, nous nous tournons vers les algorithmes CART et les modèles ensemblistes du Machine Learning comme les forêts aléatoires ou encore le XGBoost qui sont plus performants grâce aux algorithmes de Bagging et de Boosting.

Voici un tableau comparatif des modèles de ML avec des critères de comparaison. Les modèles sont notés du + (moins bonne note) au ++++ (meilleure note).

Tableau comparatif des modèles

Critères	GAM	XG BOOST	CART	Forêts aléatoires
Interprétabilité	++	+	++ (Représentation graphique)	+
Robustesse	++ (Résultats stables même sur des données non volumineuses et de qualité peu fiable)	+++	+ (Quelques variables mal renseignées ou extrêmes influencent le choix des nœuds)	+++
Pouvoir prédictif	+	++++	++	+++
Facilité de paramétrage	+++	+	++	++

Les différents critères pour obtenir le modèle de meilleure qualité et performance possibles sont maintenant présentés.

Section 3.7 : Critères de comparaison des modèles

Pour choisir le meilleur modèle de prédiction, il est important de connaître les différents critères de mesure de performance :

Indice de Gini et courbe de Lorenz : l'indice de Gini est un indice qui mesure la concentration d'objets dans une classe, c'est à dire l'impureté au sein d'une classe. Il est souvent utilisé pour mesurer les inégalités de revenus entre les pays. Il est compris entre 0 et 1. Lorsqu'il se rapproche de 0, cela signifie que la classe est plus homogène et le pouvoir prédictif du modèle est meilleur. Il se mesure par rapport à la courbe de Lorenz. Dans un contexte socioéconomique, celle-ci représente la proportion de richesse cumulée (en ordonnée) qui est détenue par un pourcentage de population la moins riche (en abscisse).

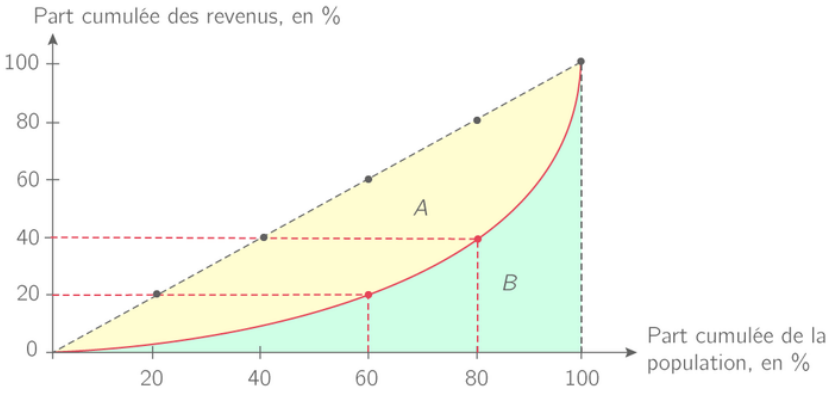


Figure 48 : courbe de Lorenz dans le contexte socio-économique

L'indice de Gini se calcule en faisant un rapport entre deux aires : Indice de Gini = A / (A+B)

La courbe de Lorenz pour notre étude est représentée de manière différente. En fait, elle est adaptée avec un graphique symétrique. Elle représente le tracé des valeurs observées en fonction de l'exposition triée selon les prédictions triées. Ainsi, la courbe exprime une mesure de segmentation du portefeuille étudié et permet donc de mesurer la qualité de notre modèle de prédiction. Si nous pensons en termes d'arrêts de travail par exemple, le graphique ci-dessous indique que 20% des personnes (ou contrats) qui ont le plus d'arrêts de travail expliquent 32% des arrêts. Autrement dit, les 20% des prédits comme les plus risqués représentent en réalité environ 32% du risque observé.

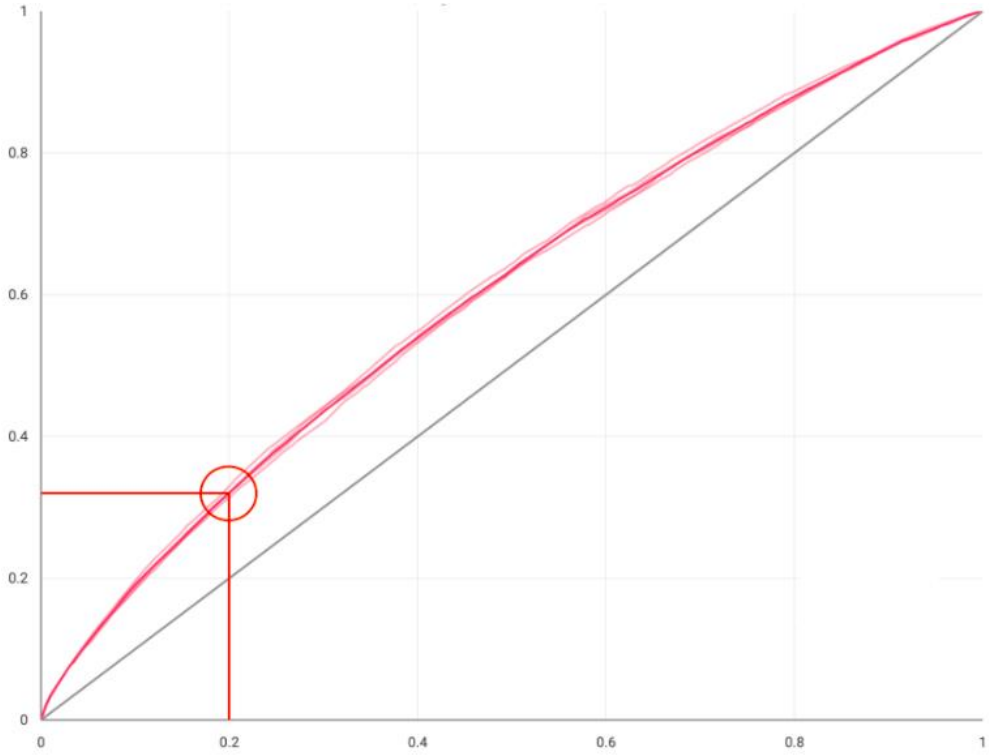


Figure 49 : exemple d'une courbe de Lorenz

La droite passant par l'origine représente le modèle aléatoire, la courbe rose foncé représente notre modèle prédictif.

RMSE (Root-mean-square-error) : il mesure l'écart qu'il y a entre les valeurs prédites du modèle et les valeurs observées. Cette différence est nommée « résidu » quand la comparaison se fait sur la base Train et « erreur de prédiction » quand elle se fait sur la base Test. Il est à noter que cette mesure capte l'effet des valeurs extrêmes et y ait sensible.

Voici la formule du RMSE :

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

Avec $\hat{y}_i$ la prédiction de la variable cible, y_i l'observation et ω_i le poids attribué à l'observation.

MSE(mean-square-error) : c'est le carré du RMSE.

MAE (Mean Absolute Error) : c'est la moyenne arithmétique des valeurs absolues des différences entre les valeurs observées et prédites.

La courbe Lift : cette courbe permet de mesurer la performance du modèle. Elle est aussi appelée courbe de gain. Elle est construite en traçant les prédictions et les observations moyennes (triées selon les prédictions) sur n groupes de l'échantillon Test de mêmes tailles. Les écarts-types des moyennes calculées sont aussi représentés dans le graphique ci-dessous.

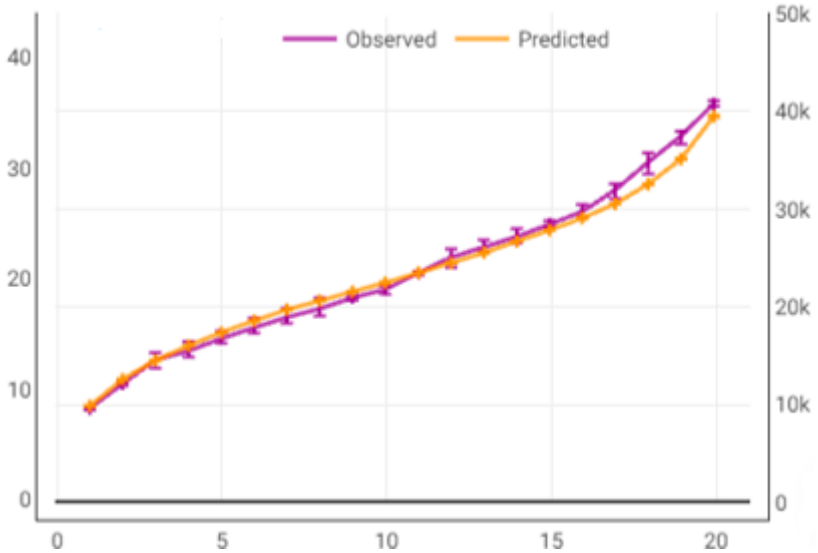


Figure 50 : exemple d'une courbe Lift

Les différentes méthodes de modélisation vont maintenant être appliquées à notre jeu de données afin de créer nos lois de fréquence et de maintien en arrêt.

En premier lieu, nous construisons une loi de maintien avec la méthode de Kaplan Meier appliquée à nos données dans l'objectif de faire une comparaison avec la loi de maintien du BCAC.

Partie 4 : comparaison de la loi de Kaplan Meier appliquée à la DSN à celle issue du BCAC

Nous comparons l'estimation de Kaplan Meier de la loi de maintien en incapacité construite sur nos données avec la loi de maintien BCAC 2013. Notre loi est créée sur les arrêts qui durent moins d'un an pour les branches « aide à domicile » et « propreté ».

Les courbes ne sont pas lissées dans ce mémoire. L'objectif est de faire une comparaison entre les lois de manière brute et pas d'utiliser ces lois - ci dans une intention de tarification.

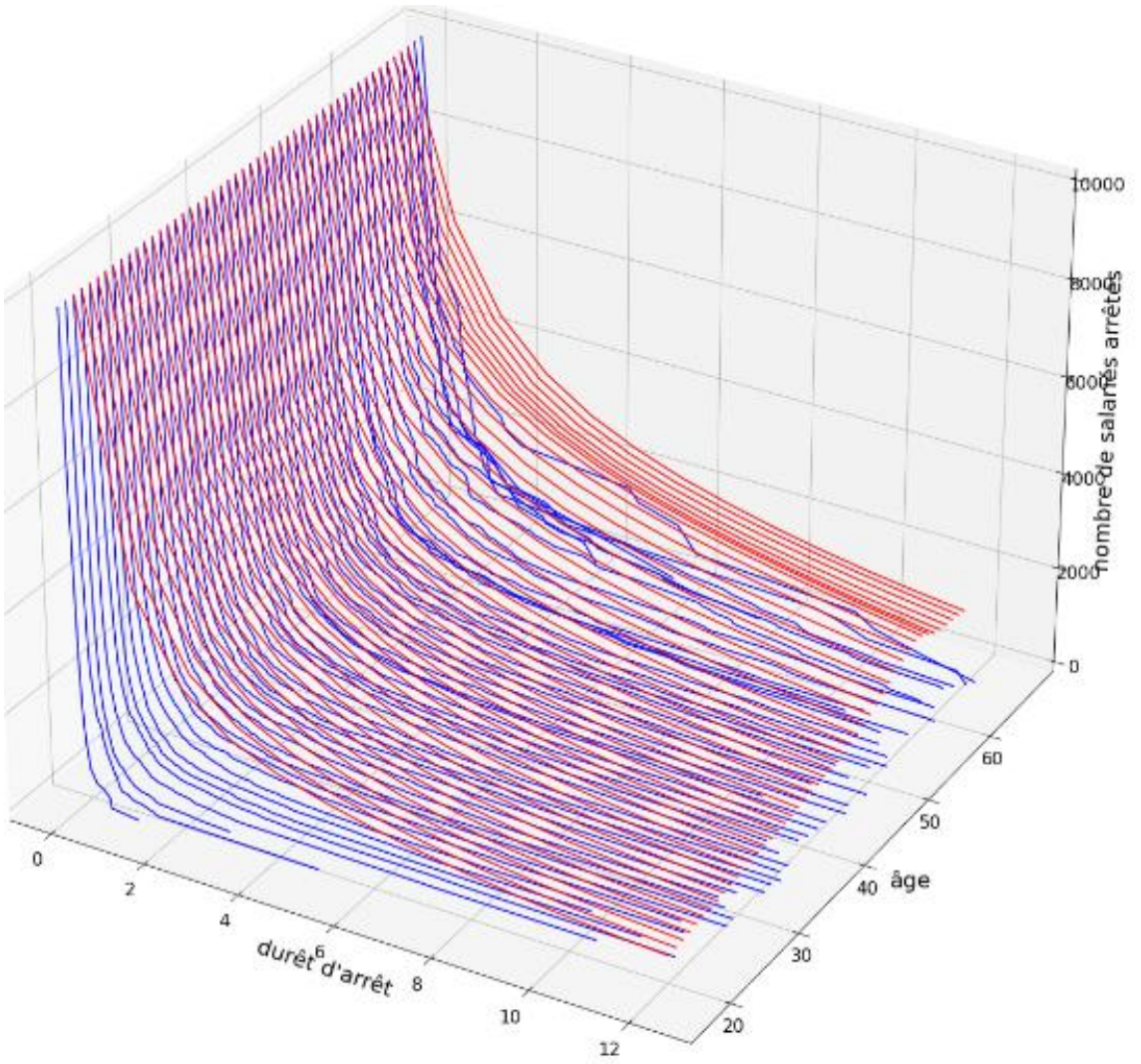


Figure 51 : comparaison des lois de Kaplan Meier constatée et du BCAC 2013

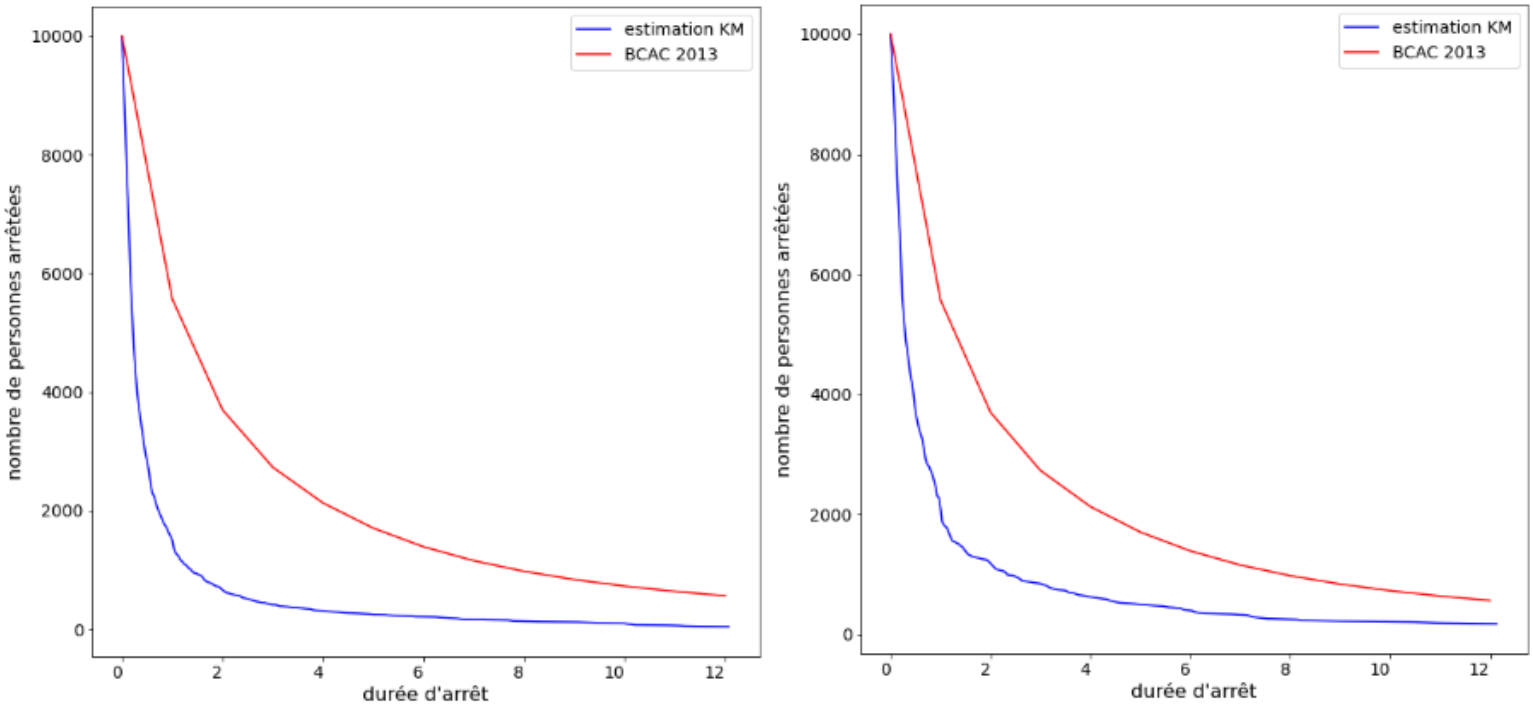


Figure 52 : zoom des courbes de KM constatée et du BCAC pour l'âge 40 ans à gauche et 55 ans à droite

Nous pouvons constater que la loi de Kaplan Meier est bien en dessous de celle du BCAC. De plus, plus l'âge augmente, plus l'estimation KM se rapproche de la loi du BCAC. Nous pouvons le voir sur les ZOOM, pour l'âge de 40 ans la courbe bleue est plus éloignée de la rouge que pour l'âge de 55 ans. Nous en déduisons que plus les salariés des branches « aide à domicile » et « propreté » sont âgés, plus leurs arrêts durent longtemps et se rapprochent des durées prédites par le BCAC.

Cependant, les tables du BCAC ont une vision uniquement sur les arrêts qui durent plus longtemps que la franchise. Pour réaliser une comparaison ayant plus de sens, les arrêts de moins de 30 jours sont éliminés dans nos données. L'estimation de KM se rapproche alors de la loi du BCAC comme le montrent les graphiques ci-dessous.

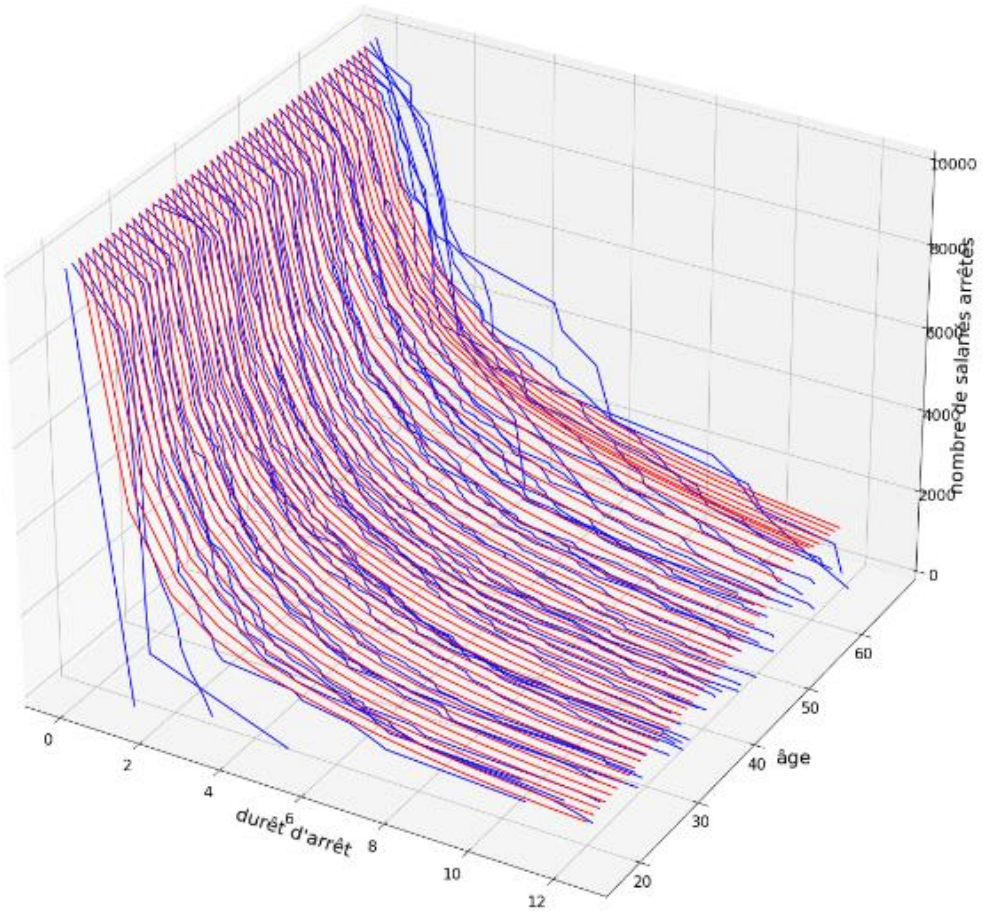


Figure 53 : comparaison des lois de Kaplan Meier constatée et du BCAC 2013 avec prise en compte de la franchise

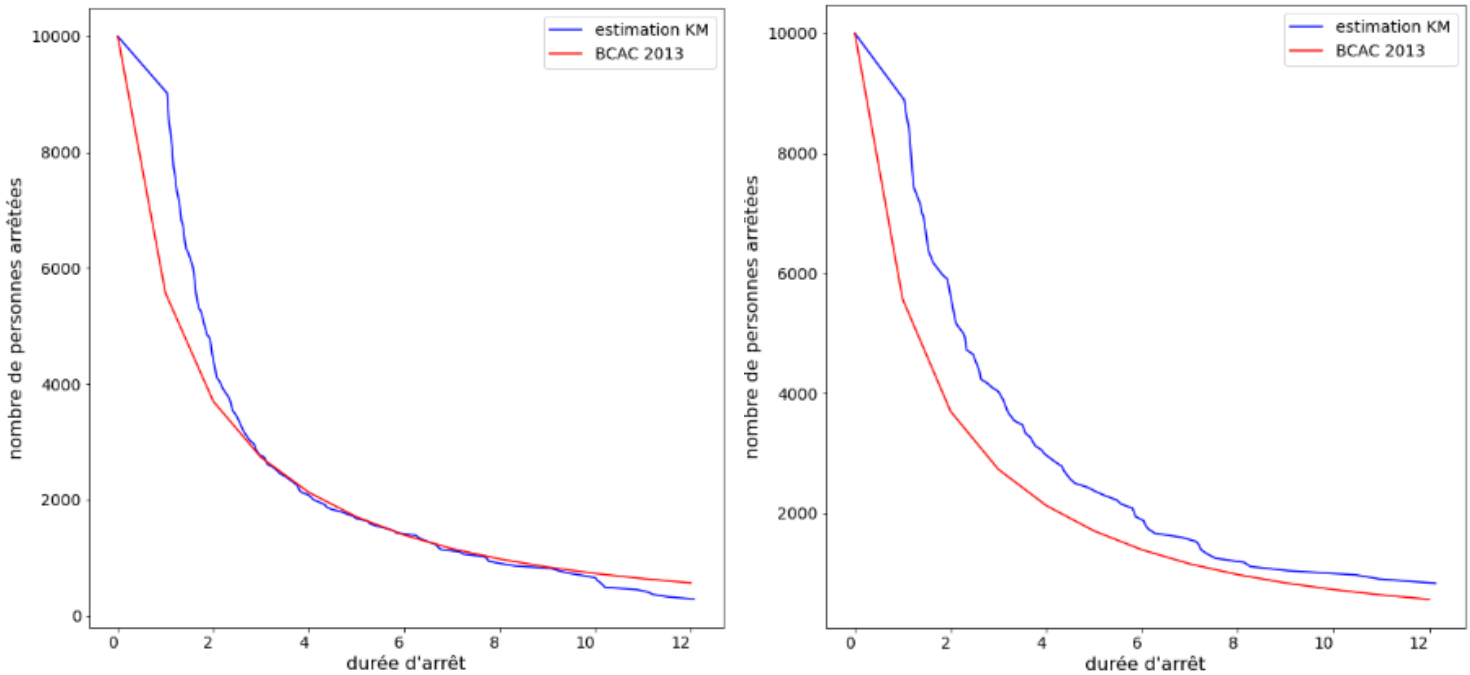


Figure 54 : zoom des courbes de KM constatée et du BCAC pour l'âge 40 ans à gauche et 55 ans à droite

Nous constatons aussi que les courbes se rapprochent plus pour les 40 ans que pour les 55 ans.
Avant trois mois d'ancienneté dans l'arrêt, l'estimation de Kaplan Meier est assez éloignée du BCAC. Cela s'explique par le fait que le BCAC n'a pas de visibilité sur les arrêts qui se terminent durant la franchise et donc estime mal la loi pour les arrêts courts. Cela justifie encore le fait d'utiliser les bases de données de la DSN qui ont une vision sur les arrêts courts.

La présence de bruitage est observée pour les âges faibles et importants (extrêmes), ceci provoquant le manque de lissage sur ces taux-ci. Il y a moins de personnes dans ces tranches-ci qui sont présentes dans la base, ce qui explique cette observation.

De cette comparaison, deux choses sont à retenir. Tout d'abord, le fait que la table du BCAC a un manque de visibilité sur les arrêts courts, ce qui confirme notre motivation à faire notre étude sur les données de la DSN. Ensuite, nous constatons que les salariés des branches « aide à domicile » et « propreté » ont des durées d'arrêts plus longues que les durées moyennes prédites par la loi du BCAC.

La loi de Kaplan Meier se base dans ce cas-là seulement sur l'âge des salariés et ne permet pas de prendre en compte les nombreuses variables de la DSN. Nous appliquons donc les méthodes de Machine Learning qui permettent de prendre en compte plusieurs variables et d'établir une loi de maintien plus précise ainsi qu'une loi de fréquence.

Partie 5 : modélisations des lois de fréquence et de maintien

Le choix a été porté pour modéliser le coût de l'arrêt de travail en deux temps :

- La modélisation de la fréquence de l'arrêt de travail pour un salarié affecté à un contrat de Prévoyance dans une entreprise. La variable à modéliser est le nombre d'arrêts survenus sur une année pour l'individu. Cette variable est pondérée par la durée de présence de l'assuré sur l'année ;
- La modélisation de la durée d'un arrêt court (entre 1 jour et 6 mois).

Le produit des deux modèles permet d'estimer la durée totale des arrêts sur un an pour un assuré d'un contrat de Prévoyance collective.

Dans cette partie, nous nous proposons de comparer les résultats de trois algorithmes du ML : CART, RF et XGB.

Section 5.1 : loi de fréquence

Dans cet exercice de modélisation :

- *Un individu* correspond à un salarié rattaché à un contrat de Prévoyance ;
- *La variable d'intérêt* correspond au nombre d'arrêts de l'individu tout au long de la durée de son exposition au risque pendant une année ;
- *La durée de l'exposition* au risque correspond au nombre de jours de présence de l'individu dans l'entreprise.

Le périmètre de l'étude est limité aux salariés des entreprises des deux branches choisies précédemment : aide à domicile et propreté.

Pour implémenter les différents algorithmes choisis, nous utilisons les fonctions disponibles à partir de la librairie Sickit-Learn du langage Python. Ces fonctions permettent de définir les échantillons d'apprentissage et de test, les mesures de performance des modèles ainsi que les hyperparamètres d'optimisation des différents algorithmes. Ci-dessous, le détail des hyperparamètres pour les algorithmes choisis :

- **CART**
  - *min_samples_split :* nombre minimum d'instances qu'un nœud doit avoir pour pouvoir être divisé ;
  - *min_sample_leaf :* nombre minimum d'instances qu'un nœud terminal doit avoir ;

- *max_depth :* profondeur maximale de l'arbre ;
- *Criterion :* critère qui mesure la qualité du « split » afin de sélectionner les variables pour la division ;
- *max_features :* nombre de variables à considérer pour choisir la variable utilisée pour le critère de division.

- **RANDOM FOREST**

  - *n_estimators :* nombre d'arbres dans la forêt ;
  - *max_features :* nombre de colonnes sélectionnées pour chaque arbre (la valeur par défaut est la racine carré du nombre de colonnes) ;
  - *Bootstrap :* paramètre pour utiliser du Bootstrap pour la construction de chaque arbre, s'il est à « False », le même échantillon est pris pour chaque arbre ;
  - *random_state :* contrôle à la fois le caractère aléatoire de l'amorçage des échantillons utilisés lors de la construction d'arbres (si Bootstrap = « True ») et l'échantillonnage des variables à prendre en compte lors de la recherche de la meilleure division à chaque nœud.

- **XGB**

  - *learning_rate :* facteur de pondération pour les corrections par de nouveaux arbres lorsqu'ils sont ajoutés au modèle pour ralentir le surapprentissage ;
  - *objective :* définit la fonction de perte à utiliser. Plus précisément, ce paramètre est la somme des fonctions Loss et de régularisation et contrôle la robustesse du modèle.

Pour chaque paramètre, plusieurs valeurs sont testées pour trouver celle qui optimise l'apprentissage de l'algorithme. Pour ce faire, la technique de la cross validation est implémentée :

- La base est divisée en 10 échantillons : 9 pour l'apprentissage et 1 pour le test ;
- A chaque itération, l'échantillon du test est fixé : il permet de tester les performances du modèle initialisé avec un set fixe d'hyper paramètres ;
- Le set d'hyper paramètres qui donne la meilleure précision du modèle est retenu pour l'algorithme.

Opérationnellement, cette étape est automatisée dans Python avec la fonction « *GridSearch* ».

Voici les paramètres retenus après cette optimisation :

Critères	CART	RANDOM FOREST	XGB
max_depth	10	12	10
min_sample_leaf	50	30	30
min_samples_split	50	50	50
criterion		« MSE »	« MSE »
max_features		« Sqrt »	« Sqrt »
n_estimators		100	100
bootstrap		« True »	« True »
random_state		42	42
learning_rate			0,1
objective			« Count :Poisson »

Après la création des modèles, les résultats sont présentés.

- Temps d'exécution du programme

	CART	Random Forest	XGB
Coût en calcul (en secondes)	2,22 s.	141,82 s.	222,29 s.

Le Random Forest est près de 63 fois plus coûteux en calcul que le modèle CART pour la modélisation de la fréquence. Le XGB, lui, est 1,5 fois plus coûteux que le Random Forest.

- Indicateurs de performance globaux

	CART	Random Forest	XGB
RMSE	6,88	6,55	6,46
MAE	1,76	1,73	1,58
MSE	43,40	42,94	41,82
Gini	0,22	0,26	0,28

Le modèle XGB affiche l'erreur la plus faible et le Gini le plus important. Il a donc la meilleure performance. Nous savons aussi que la fréquence moyenne observée de l'échantillon est de 0,20 avec un écart-type de 0,63 et nous pouvons comparer au MAE pour le XGB qui est de 1,58. L'erreur absolue moyenne de prédiction est donc assez importante et provient probablement de la présence de fréquences extrêmes dans la base.

En plus des mesures des erreurs globales, les modèles sont comparés en utilisant les courbes de lift et de Lorenz.

Pour construire la courbe Lift, nous procédons comme suit :

- Trier les individus par niveaux de prédictions croissants ;
- La population est ensuite séparée en 20 échantillons de taille égale ;
- Pour chaque échantillon, deux valeurs sont calculées : la moyenne des prédictions et la moyenne des observations.

La courbe est ainsi tracée pour pouvoir comparer les prédictions moyennes aux valeurs réelles pour chaque échantillon.

Les courbes de Lorenz sont construites en triant les observations et les expositions selon les prédictions décroissantes et en traçant les expositions cumulées observées et normées en fonction des prédictions (nombre d'arrêts prédit) cumulées triées normées.

Ces résultats sont présentés pour les arbres CART, le Random Forest et le XGBoost.

5.1.1 Détails des résultats du modèle CART

La visualisation de l'arbre modélisé par CART permet d'identifier les variables les plus discriminantes dans l'exercice de régression. Les premiers critères de division retenus sont le statut conventionnel, la rémunération

mensuelle moyenne, l'âge du salarié, la région de l'entreprise, le montant de rémunération annuel du salarié et le libellé APE.

Ces variables reflètent à la fois l'environnement du travail du salarié (la convention, le code APE, la région et la rémunération moyenne au sein de l'entreprise) ainsi que les caractéristiques du salarié (l'âge et sa rémunération annuelle).

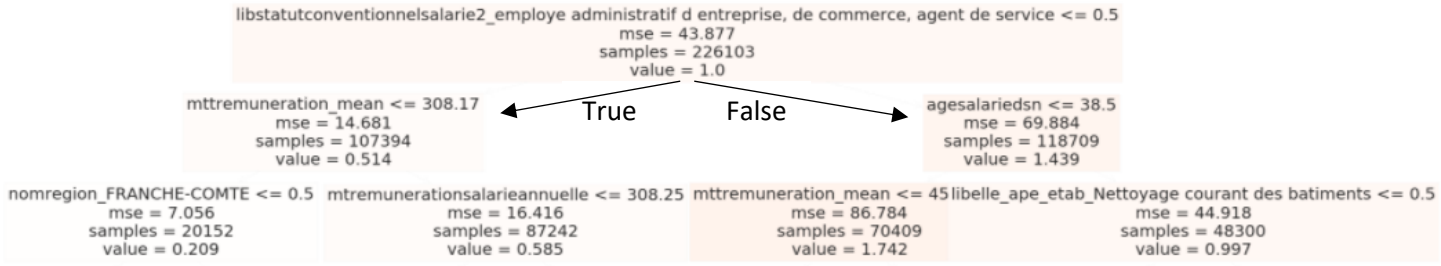


Figure 55 : extrait de l'arbre CART pour le modèle de fréquence

L'arbre indique à chaque nœud le critère mse, la taille de l'échantillon (samples) et la fréquence d'arrêts prédite (value).

- Courbe Lift

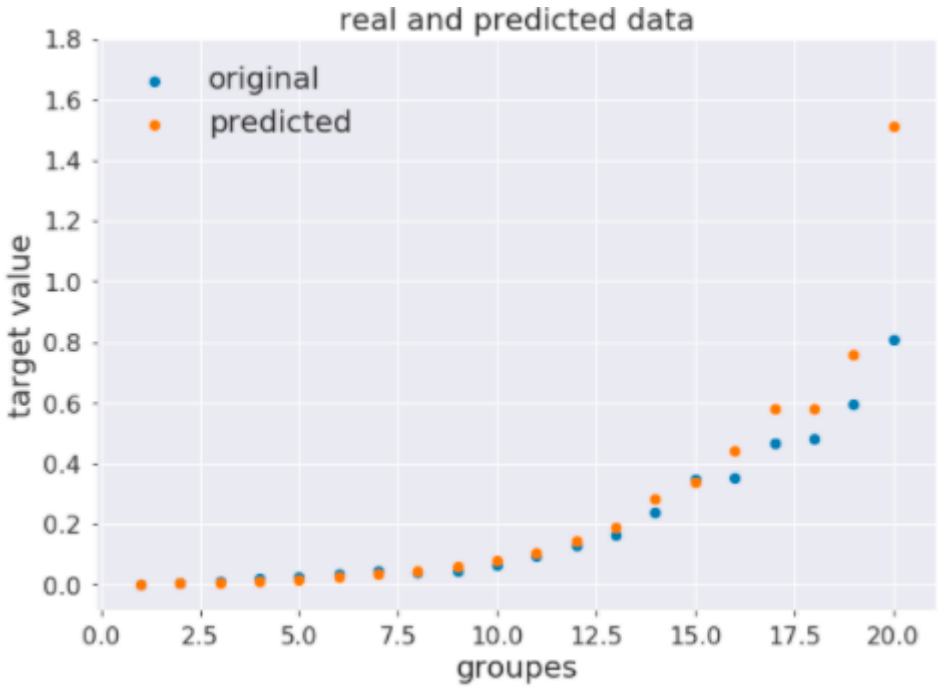


Figure 56 : courbe Lift pour le modèle de fréquence CART

La précision du modèle se dégrade sur les observations les plus élevées :

- Le modèle surestime les observations à partir de 0,4 arrêt ;
- A partir de la moyenne observée de 0,8 arrêt, l'écart devient plus important et atteint 0,7.

➔ Le modèle CART est moins précis sur les valeurs extrêmes élevées.

5.1.2 Détails des résultats du modèle Random Forest

- Courbe Lift

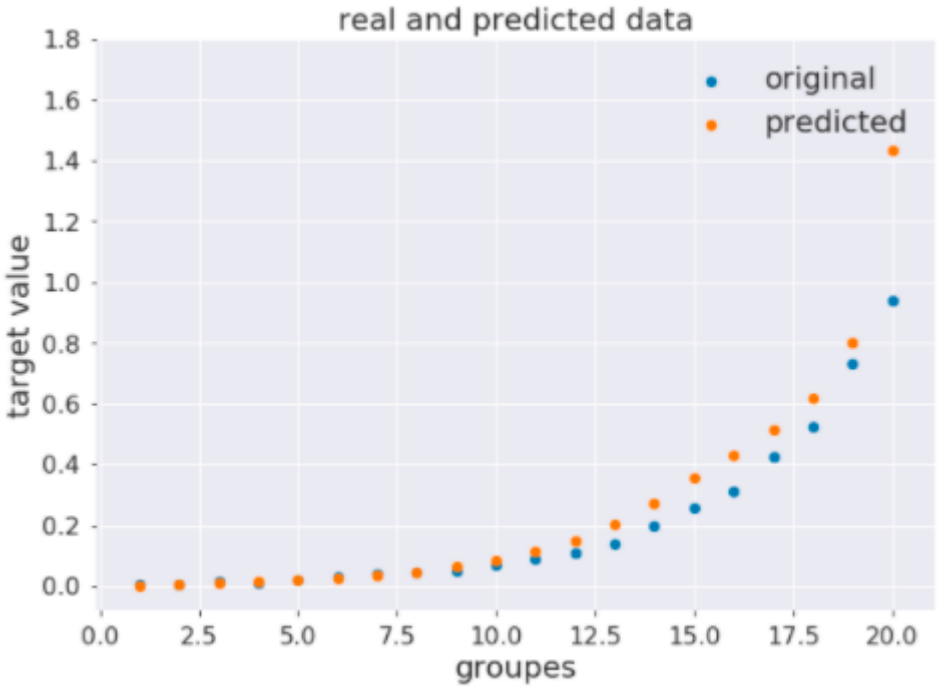


Figure 57 : courbe Lift pour le modèle de fréquence Random Forest

Par rapport à la courbe Lift précédente, une amélioration est observée pour les observations moyennes supérieures à 0,5 arrêt. Les écarts sont amoindris entre l'observé et le prédit, et le dernier groupe atteint maintenant un écart de 0,4 arrêt. Néanmoins, les écarts sont plus importants entre 0,2 et 0,4 arrêt prédits (nombres moyens).

Cela signifie que le Random Forest a amélioré la prédiction des nombres d'arrêts importants mais pas forcément des niveaux d'arrêts moins élevés.

5.1.3 Détails des résultats du modèle XGB

- Courbe Lift

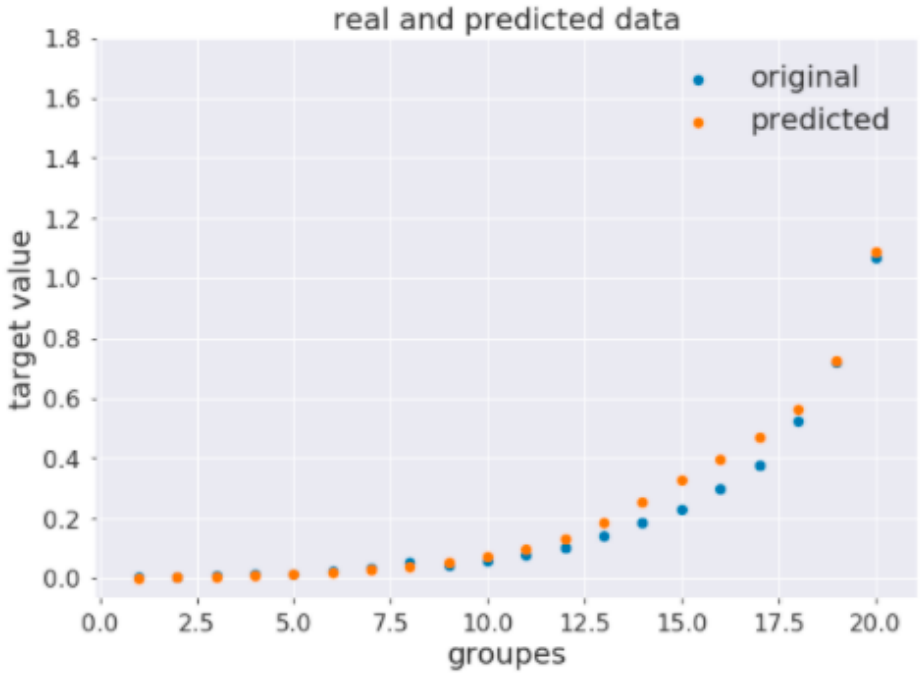


Figure 58 : courbe Lift pour le modèle de fréquence XGB

Le XGB apporte une nette amélioration de la performance par rapport aux modèles précédents. Nous constatons que les écarts entre l'observé et le prédit pour les grands nombres d'arrêts sont plus faibles que sur les courbes Lift précédentes. Pour les deux derniers groupes d'arrêts observés (après 0,6 arrêt), l'amélioration est considérable. Le XGB permet donc d'améliorer la prédiction surtout sur les nombres d'arrêts supérieurs à 1.

Et voici les courbes de Lorenz qui nous permettent de mesurer la qualité de segmentation du portefeuille.

- Courbe de Lorenz

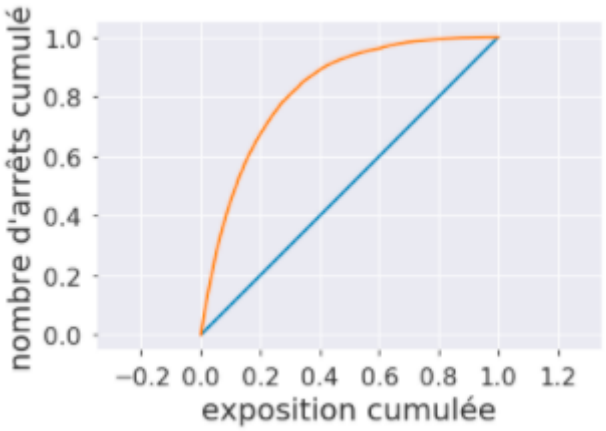


Figure 59 : courbe de Lorenz pour le modèle de fréquence XGB

Cette courbe de Lorenz est comparée à celles des modèles CART et Random Forest.

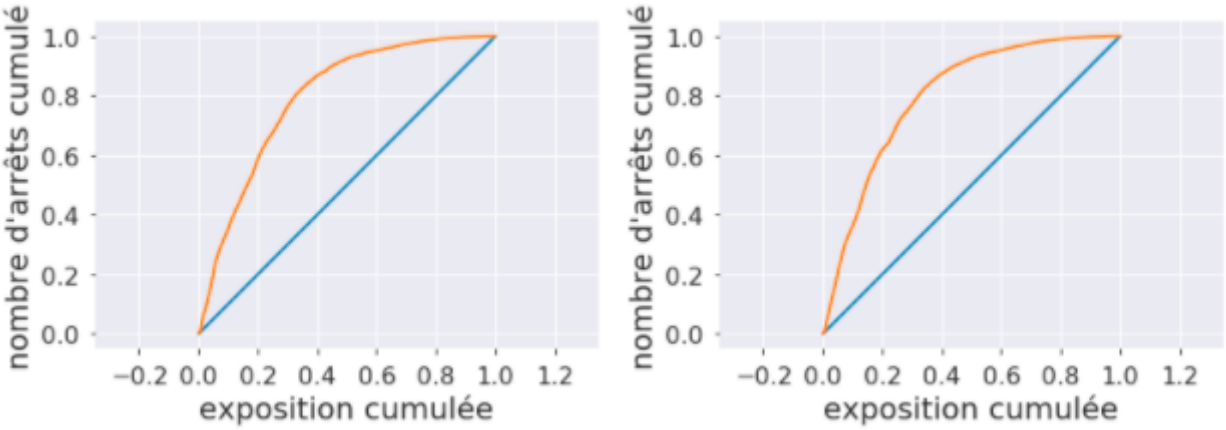


Figure 60 : courbes de Lorenz pour les modèles de fréquence CART (gauche) et Random Forest (droite)

Nous observons qu'elles deviennent de plus en plus concaves, et le Gini devient de plus en plus important lorsqu'on passe du CART, au RF puis au XGB. Si la construction de la courbe de Lorenz est changée en triant les expositions par les observations et non par les prédictions toujours dans l'ordre décroissant, nous obtenons :

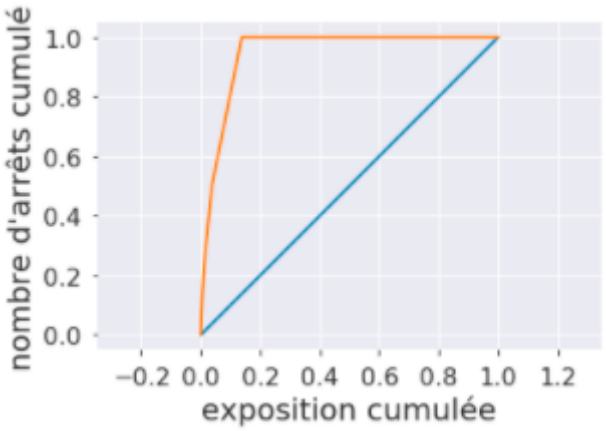


Figure 61 : courbe de Lorenz théorique pour le modèle de fréquence XGB

Cette courbe montre donc que 15% environ des contrats présentent en réalité des arrêts et la courbe de Lorenz du XGB précédente montre qu'avec seulement 20% des contrats, nous détenons 70% des sinistres, ce qui est une bonne segmentation. Plus les courbes sont proches de l'observé (la courbe de Lorenz avec le tri sur les observés), meilleure la prédiction est. La courbe du XGB est celle qui en est la plus proche.

Le modèle XGB est donc le meilleur modèle parmi les trois présentés. Voici les variables qui sont sélectionnées par le modèle par ordre d'importance.

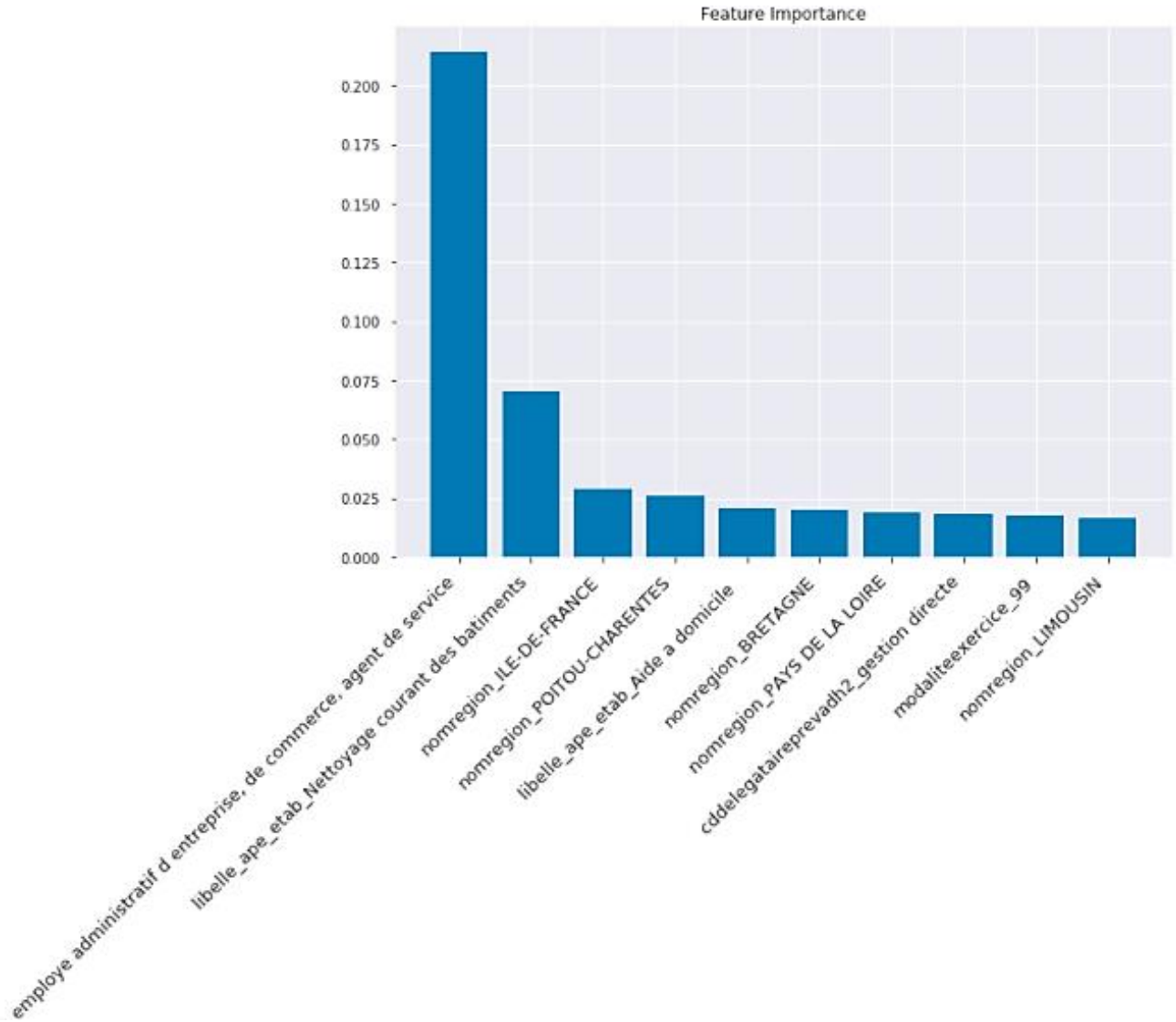


Figure 62 : importances des variables sélectionnées pour le modèle de fréquence XGB

Les variables les plus contributives sont :

- Le statut conventionnel du salarié : employé administratif d'entreprise, de commerce et agents de service ;
- L'APE de l'établissement (nettoyage courant des bâtiments et aide à domicile) : l'appartenance à une entreprise de la branche « aide à domicile » discrimine le nombre d'arrêts observé ;
- L'emplacement géographique de l'entreprise est aussi discriminant dans le modèle ;
- Le type de gestion directe ou déléguée ressort dans l'analyse. Cette variable peut être liée à la taille de l'entreprise ou à la spécificité du portefeuille.

Comme vu précédemment, certaines variables sont en commun avec celles retenues dans le modèle CART. Ce classement peut être comparé à ceux pour le CART et le Random Forest en **annexe 6**.

Section 5.2 : loi de maintien

La loi de maintien en arrêt permet de modéliser la durée probable d'un arrêt conditionnellement à son occurrence. Pour cet exercice :

- L'individu est un arrêt déclaré dans la DSN ;
- La variable d'intérêt est la durée de l'arrêt, pour les arrêts de durée allant de 1 jour à 6 mois.

Ci-dessous les paramètres optimaux obtenus grâce au *Gridsearch* et implémentés dans nos modèles.

Critères	CART	RANDOM FOREST	XGB
max_depth	12	12	10
min_sample_leaf	60	30	40
min_samples_split	40	50	50
criterion		« MSE »	« MSE »
max_features		« Sqrt »	« Sqrt »
n_estimators		100	100
bootstrap		« True »	« True »
random_state		42	42
learning_rate			0,1
objective			« reg :Gamma »

Nous présentons maintenant les résultats que nous obtenons en paramétrant les modèles ainsi.

➢ Temps d'exécution du programme

	CART	Random Forest	XGB
Coût en calcul (en secondes)	0,46 s.	3,80 s.	37,45 s.

Le Random Forest prend environ 8 fois plus de temps à s'exécuter que le modèle CART pour le modèle de durée. Et le XGB est près de 9 fois plus coûteux en calcul que le Random Forest.

➢ Indicateurs de performance globaux

Critères	CART	Random Forest	XGB
RMSE	24,20	23,77	23,70
MAE	14,21	13,90	13,80
MSE	585,64	565,10	561,71
Gini	0,08	0,10	0,11

Nous remarquons ici que les indicateurs de performance sont meilleurs pour le XGBOOST que pour les autres modèles avec un Gini plus fort et des RMSE, MAE et MSE plus faibles.

La durée moyenne observée de l'échantillon est calculée : 15,23 avec un écart-type de 24,40. Ces observations peuvent être comparées au MAE du XGB de 13,80 et cela met en avant l'importance de l'erreur absolue moyenne.

Comme pour la loi de fréquence, les courbes de Lorenz et les courbes Lift sont comparées afin d'aboutir au choix du modèle.

5.2.1 Détails des résultats du modèle CART

Voici un extrait de l'arbre CART obtenu. Nous pouvons constater que les premiers critères de division sont la nature du contrat (CDI ou CDD), l'ancienneté du salarié dans l'entreprise, le montant de rémunération annuel, le secteur (lourd, moyen, léger) de l'entreprise et l'âge du salarié.

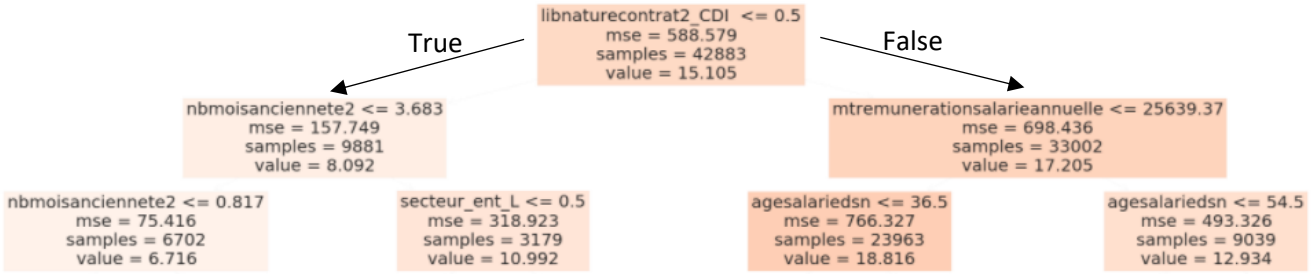


Figure 63 : extrait de l'arbre CART pour le modèle de durée

- Courbe Lift

Figure 64 : courbe Lift pour le modèle de durée CART

Nous constatons que le modèle CART ne prédit pas bien les durées d'arrêts car de nombreux écarts entre le prédit et l'observé sont observés. Cet écart est particulièrement important au niveau des durées longues (pour les durées observées de plus de 24 jours) où il peut atteindre une dizaine de jours de différence en moyenne entre l'observé et le prédit. D'autre part, nous remarquons que les prédictions sont moins élevées que les observations avant le 11ème groupe de durée moyenne observée égale à 19 jours, et c'est l'inverse après le 18ème groupe. Le modèle surestime donc les durées les plus importantes et sous-estime les plus faibles.

5.2.2 Détails des résultats du modèle Random Forest

- Courbe lift

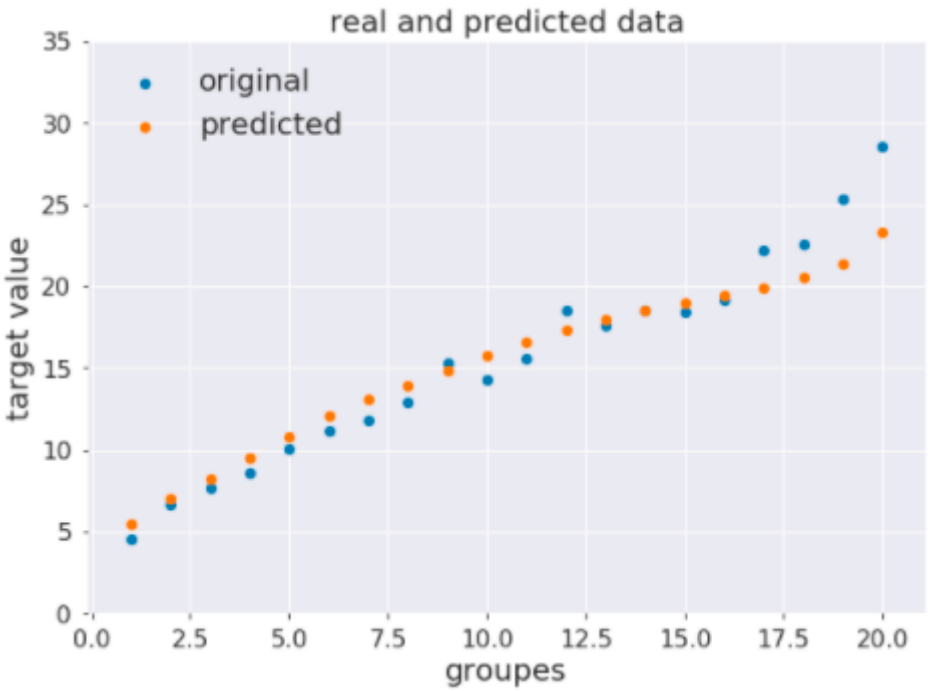


Figure 65 : courbe Lift pour le modèle de durée RF

Nous constatons une amélioration des prédictions par les forêts aléatoires. Les écarts entre les durées observées et prédites persistent toujours mais se sont globalement amoindris surtout sur les extrêmes où l'algorithme CART montrait de grandes différences entre le prédit et l'observé allant jusqu'à 10 jours en moyenne alors qu'ici l'écart va jusqu'à 5 jours.

5.2.3 Détails des résultats du modèle XGB

- Courbe Lift

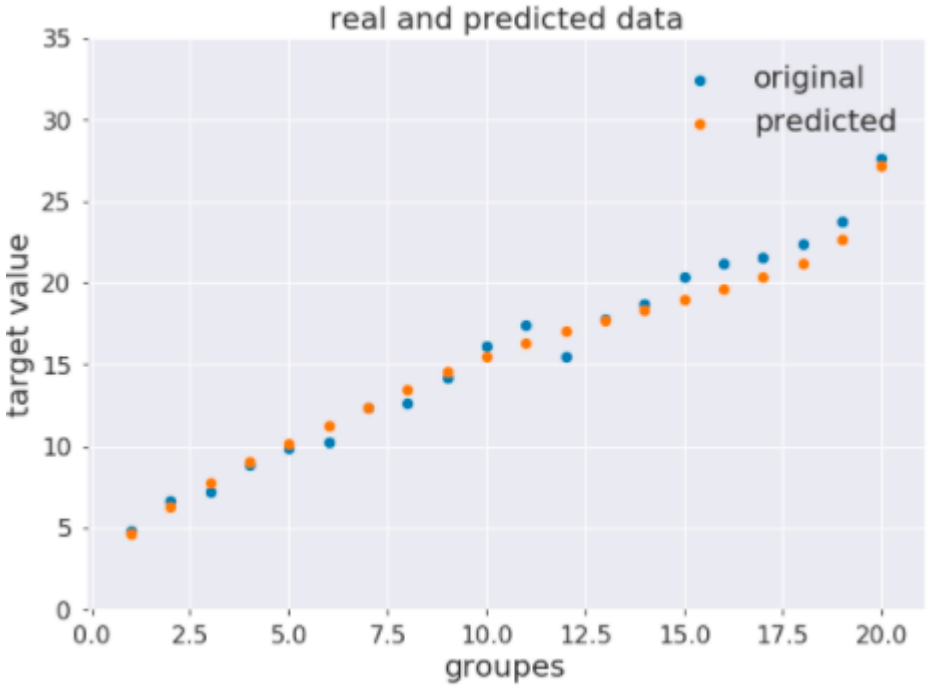


Figure 66 : courbe Lift pour le modèle de durée XGB

Les performances sont encore améliorées avec le XGB sur les durées d'arrêts importantes. Nous observons que pour les durées observées de 20 jours et plus, la prédiction moyenne est très proche de l'observation moyenne par rapport à la courbe Lift précédente.

- Courbe de Lorenz

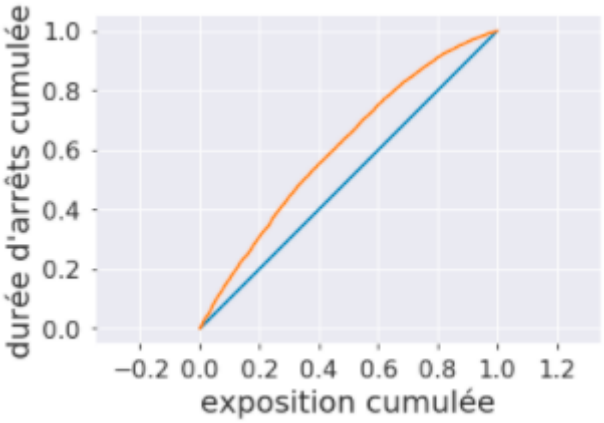


Figure 67 : courbe de Lorenz pour le modèle de durée XGB

Cette courbe de Lorenz est comparée à celles des modèles CART et Random Forest.

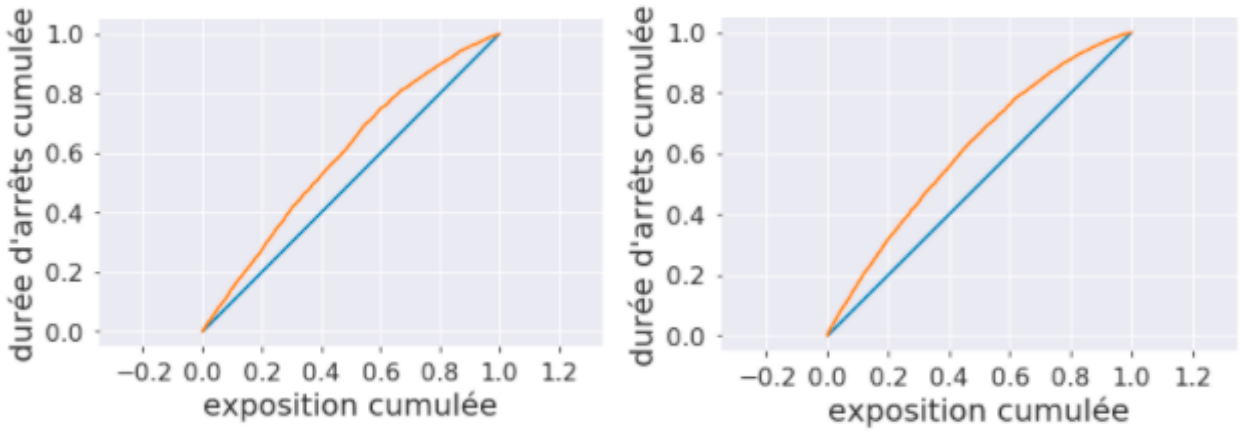


Figure 68 : courbes de Lorenz pour les modèles de durée CART (gauche) et Random Forest (droite)

Nous observons que la courbe de Lorenz devient de plus en plus concave lorsque nous passons du CART, au RF puis au XGB.
Si la construction de la courbe de Lorenz est modifiée en triant les expositions par les observations et non par les prédictions toujours dans l'ordre décroissant, nous obtenons :

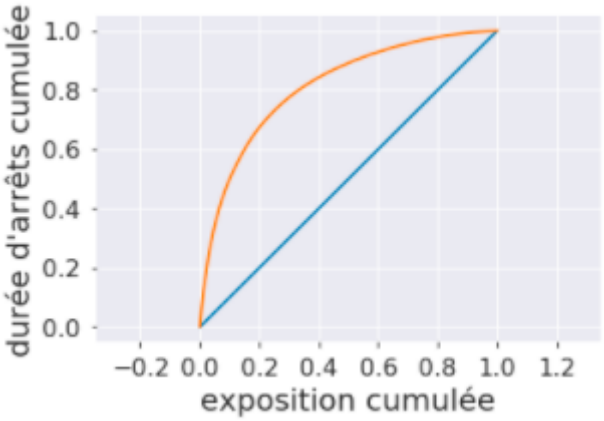


Figure 69 : courbe de Lorenz théorique pour le modèle de durée

Le modèle avec une courbe de Lorenz parfaite donnerait la même courbe de Lorenz que celle-ci. Cette courbe est interprétable en remarquant qu'environ 20% des contrats détiennent 70% des durées d'arrêts les plus longues. Autrement dit, 20% des contrats détiennent 70% de jours d'arrêts observés. Tandis que pour le XGB, 20% des contrats en détiennent seulement 30%.

Nous pouvons enfin regarder la contribution des variables au modèle de durée XGB :

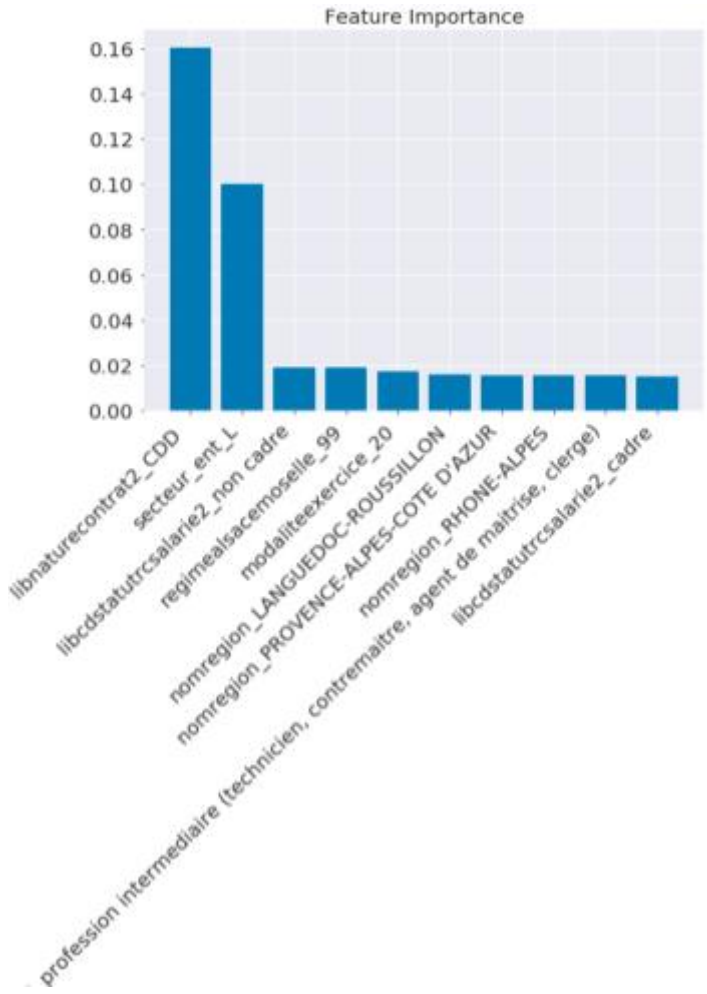


Figure 70 : importances des variables sélectionnées pour le modèle de durée XGB.

Les variables les plus importantes sont la nature du contrat, le secteur (lourd, moyen, léger) de l'entreprise, le statut du salarié, son régime, sa modalité d'exercice (temps partiel ou temps plein), la région de l'entreprise et le statut conventionnel du salarié. Ce classement peut être comparé à ceux pour le CART et le Random Forest en **annexe 6**.

A présent, nous construisons des modèles GLM. Ils sont traités à part car ils ne sont pas élaborés avec le même outil (Python). L'utilisation de cet outil de tarification externe est un défi pour l'entreprise et permet de répondre au besoin métier.

Partie 6 : évaluation d'un outil de tarification externe

Les algorithmes de Machine Learning ont pu être exploités pour construire nos modélisations en partant de données massives. Dans cette partie, est présentée un outil de tarification externe qui présente de nombreux avantages. Il permet de prendre en compte plus de données (variables), d'automatiser la gestion de celles-ci, de tester rapidement plusieurs modèles, d'automatiser les « Grid Search » et d'adapter facilement les modélisations en fonction de l'évolution des données.

Grâce à l'agilité de cet outil là et à sa vitesse de computation des calculs, les modèles lancés sur cet outil contiennent plus de variables explicatives et sont par construction différents de ceux optimisés sur notre machine en local.

Nous présenterons dans la suite les résultats des modélisations des lois de fréquence et de maintien issus de cet outil. Ces modèles vont ensuite être agrégés par une multiplication afin d'obtenir la loi finale prédisant le nombre de jours moyen d'arrêts sur un an qui nous servira à la tarification.

Section 6.1 : modèle de fréquence

Le modèle de fréquence est construit avec l'hypothèse que la variable cible suit une loi de Poisson.

Contrairement aux modèles précédents, les variables ne sont pas présélectionnées pour le modèle. Toutes sont intégrées dans l'outil, y compris celles qui contiennent trop de modalités et celles qui ont des valeurs manquantes (éliminées pour les autres modèles) car l'outil a la capacité de les gérer. Le modèle GLM fait une sélection des variables (pour un nombre choisi allant de 0 à 30 dans notre cas) qui ont les spreads les plus forts.

L'outil laisse la possibilité de créer des modélisations GAM plus générales que les GLM en ajoutant par exemple des interactions entres variables dans le modèle (exemple : sexe × âge). En testant cela, nous constatons que l'intégration de celles-ci n'améliore pas dans notre cas la performance de nos modèles. Ce sont donc des modélisations GLM qui sont retenues.

Choix du modèle

Le schéma ci-dessous met en évidence les différents modèles GLM générés par des Grid Search.

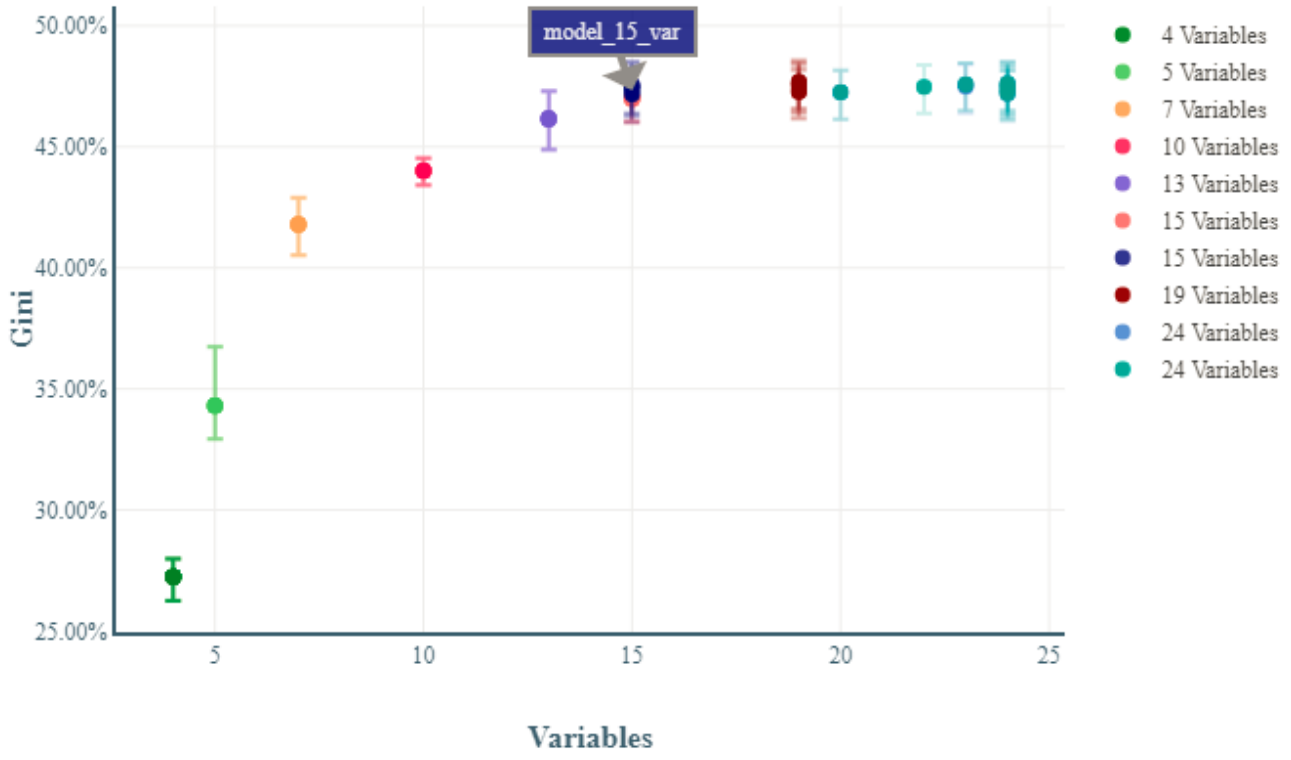


Figure 71 : Gini des modèles en fonction du nombre de variables

Dans la figure ci-dessus, chaque couleur représente une série de modèles construits avec un nombre de variables indiqué en abscisse. Pour chaque série représentée par une même couleur, les modèles diffèrent selon leur niveau de régularisation.

Le paramètre de régularisation évalue la sensibilité aux signaux faibles : un modèle régularisé suivra principalement les grandes tendances des données mais ne capturera pas toutes les variations, alors qu'un modèle non régularisé peut capturer des variations plus subtiles, parfois au détriment de la robustesse. Les coefficients du GLM représentant les effets des modalités des variables sur la réponse sont alors régularisés. Si la variable est catégorielle, un test statistique est mis en place dont l'hypothèse nulle est "le coefficient associé à cette modalité est 0". Si celle-ci est ordinale, l'hypothèse est "le coefficient associé à cette modalité est égal à son voisin". La régularisation correspond ainsi à la sévérité du test 1 - α. Plus le test est sévère, plus les coefficients seront groupés soit à zéro pour une variable catégorielle soit entre voisins pour une variable ordinale.

Le modèle avec le meilleur Gini est choisi. Le modèle encadré en bleu qui est construit avec 15 variables a un bon Gini (environ 47%) comparé aux autres modèles. Mais, en ajoutant les variables contenant les informations géographiques, nous obtenons un modèle qui est légèrement meilleur avec cette fois-ci 19 variables. Cependant, il faut faire le meilleur compromis entre le nombre de variables et la performance du modèle afin de ne pas le complexifier donc c'est le modèle à 15 variables qui est sélectionné.

Grâce au modèle comportant les informations sur la géographie, nous pouvons observer l'influence du lieu de travail sur la fréquence d'arrêt.
Voici ci-dessous la carte géographique permettant de voir les régions des entreprises qui détiennent le plus d'arrêts et celles qui en ont le moins. Nous constatons donc que pour les branches « aide à domicile » et « propreté », il y a plus d'arrêts dans les régions d'Alsace, Bretagne et Corse. En Ile de France et en Provence-

Alpes-Côte d'Azur, il y a particulièrement peu d'arrêts par rapport aux autres régions. Cela correspond à ce qui est observé dans les statistiques sur la région d'arrêt.

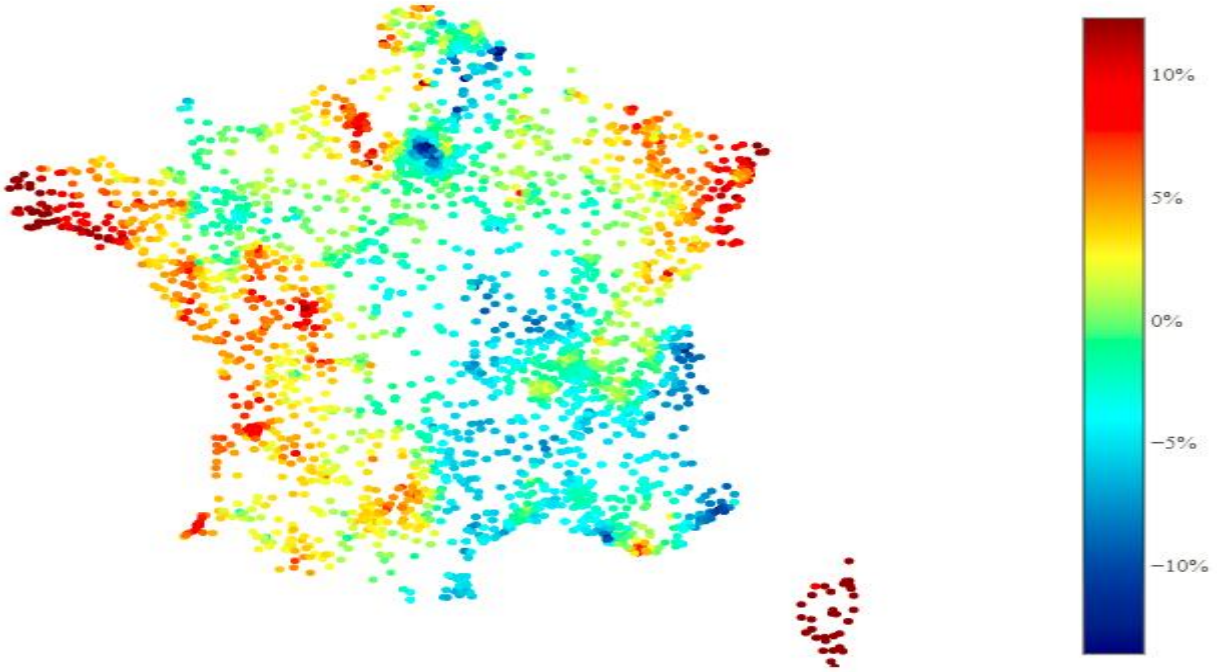


Figure 72 : cartographie des fréquences d'arrêts

Nous revenons à notre modèle à 15 variables sélectionné et elles sont présentées selon leur ordre d'importance.

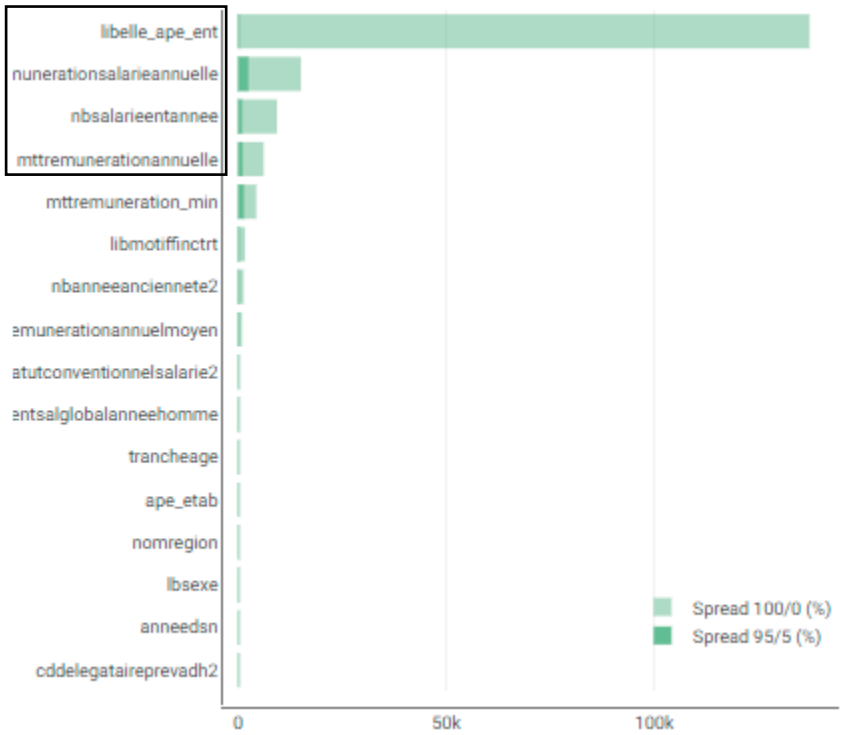


Figure 73 : classification des variables selon leur importance pour le modèle de fréquence

Interprétation du graphique : le spread mesure l'importance des variables dans le modèle. C'est le ratio entre le coefficient (effet sur la fréquence d'arrêt par rapport à la moyenne) de la variable le plus important et le plus faible. Le spread 100 est le spread mesuré sur toutes les modalités de la variable. Le spread 95 est la mesure du spread sur les 5 % des catégories qui ont les coefficients les plus faibles et sur les 5% qui ont les coefficients les plus forts.
Les variables les plus importantes (encadrées) pour le modèle constituent nos variables de tarification.

Les effets de quelques variables sur la fréquence d'arrêt sont détaillés à l'aide du tracé des coefficients accordés à chacune des modalités de la variable concernée.
Un coefficient représente l'effet que la modalité d'une variable a sur la variable cible. Il provient en fait de la formule du GLM utilisée :

$$\hat{Y} = \mathbb{E}\,[Y \mid X = \mathrm{x}] = g^{-1}\left(\sum_{\substack{i\in[\![1,p]\!] \\ j\in[\![1,q_i]\!]}} x_{i,j} * \hat{\beta}_{i,j}\right)$$

- $\hat{Y}$: la prédiction du modèle ;
- X : le vecteur des variables $X_{i\in[\![1,p]\!]}$ et x son observation ;
- g : la fonction de lien log ;
- p : le nombre de variables ;
- q_i : le nombre de modalités associé à la variable X_i ;
- $x_{i,j}$: variable n°i pour sa modalité n°j encodée en binaire : 1 si X contient cette réalisation, 0 sinon ;
- $\hat{\beta}_{i,j}$: le coefficient estimée pour la modalité n°j de la variable X_i.

En fait, g étant une fonction logarithme, en l'inversant nous obtenons un coefficient associé à une modalité décrit par e^{β}, avec β le paramètre estimé par la méthode du maximum de vraisemblance.

Voici les effets du montant de revenu annuel du salarié sur la fréquence d'arrêt.

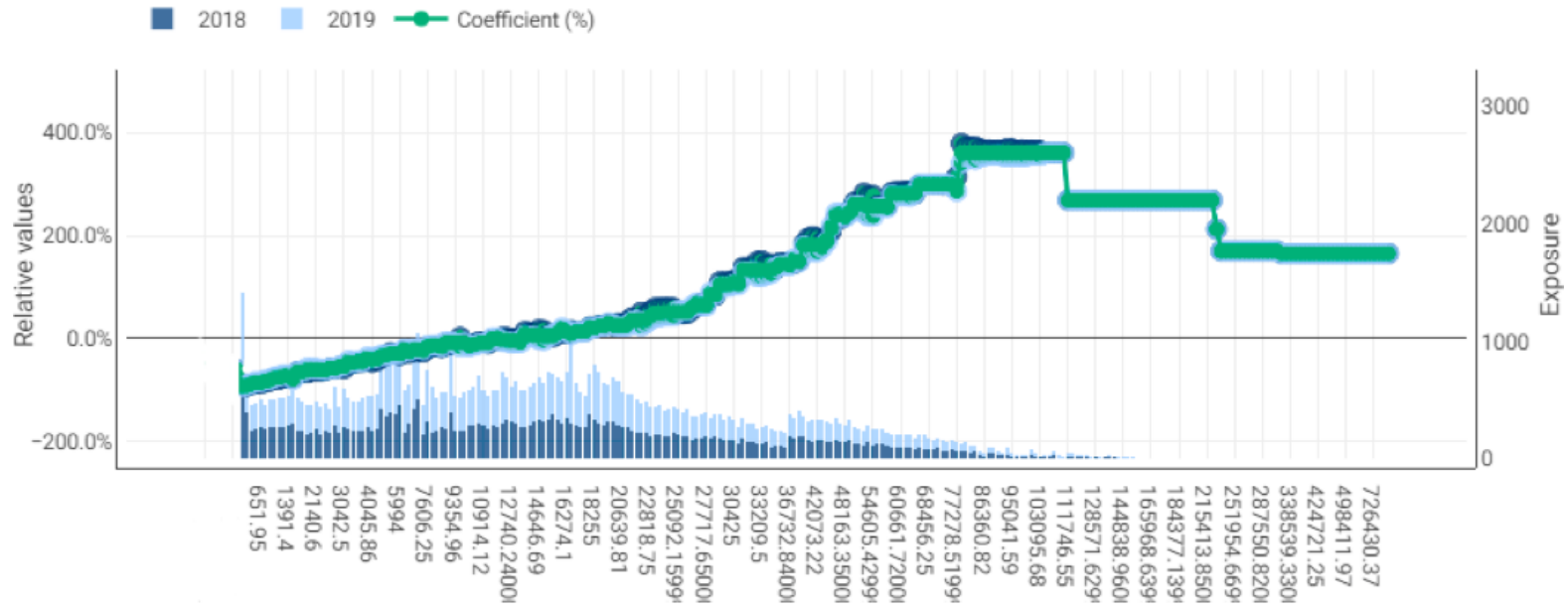


Figure 74 : effet du montant de revenu annuel du salarié sur la fréquence d'arrêts

La courbe des coefficients est croissante jusqu'à 111 000 € puis décroissante. Nous pouvons constater que les salariés qui ont plus de 30 000 € de revenu annuel expliquent le plus l'arrêt de travail. Ce sont les salariés les plus riches qui ont une fréquence d'arrêts importante bien que l'effet du salaire sur la fréquence diminue après un seuil de richesse de 111 000 €.

Les effets du nombre de salariés par entreprise sur la fréquence d'arrêt sont maintenant présentés.

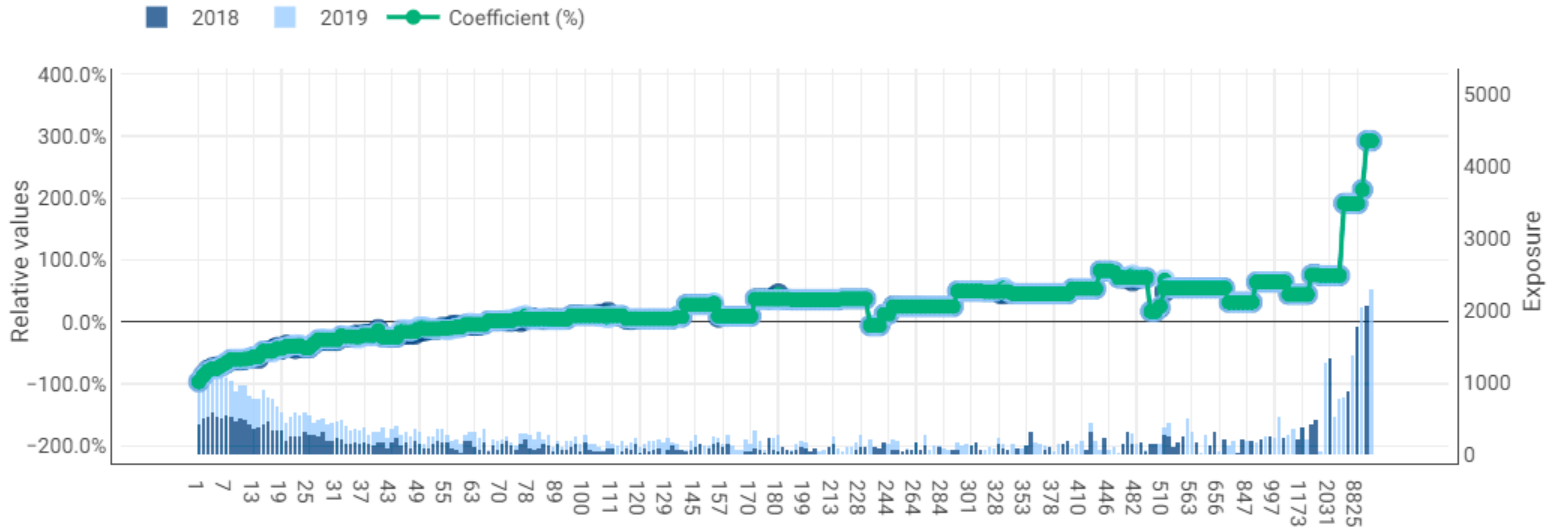


Figure 75 : effet du nombre de salariés par entreprise sur la fréquence d'arrêts

Nous constatons que les petites entreprises influencent négativement le nombre d'arrêts tandis que les entreprises de taille moyenne influencent de manière positive le nombre d'arrêts. Il y a moins d'arrêts dans les petites entreprises que dans les moyennes, c'est ce qui a été vu dans les statistiques descriptives.

A présent, nous regardons les effets des âges sur la fréquence d'arrêt.

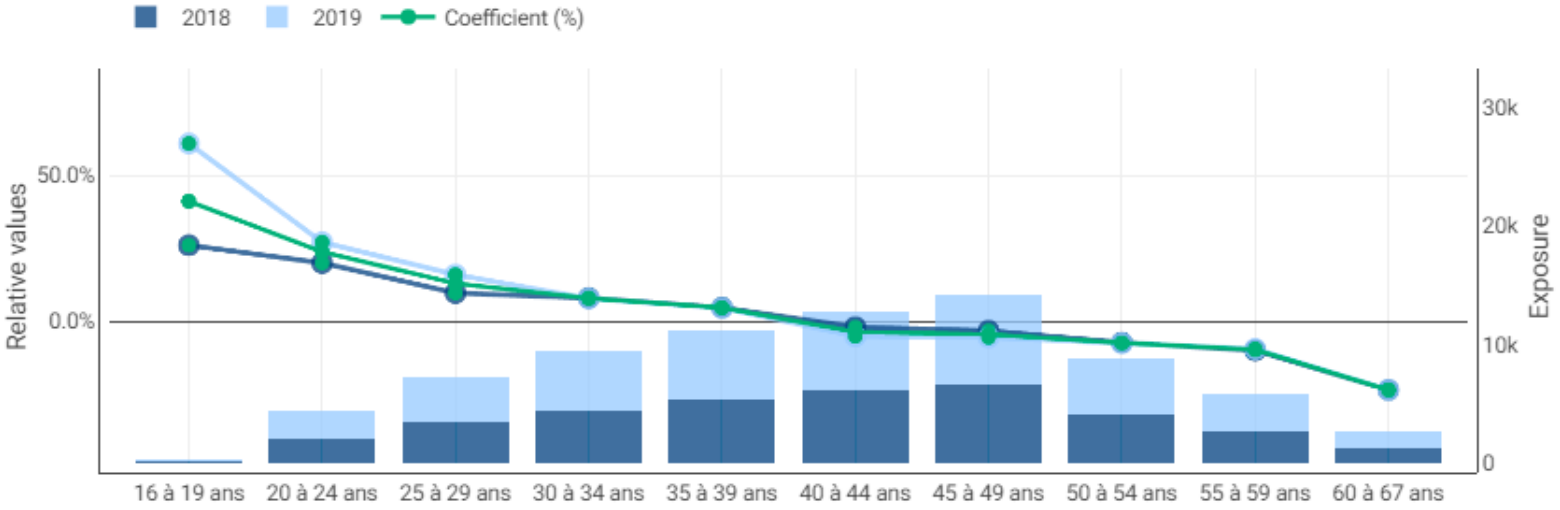


Figure 76 : effet des tranches d'âge sur la fréquence d'arrêts

Le graphique nous fait constater que ce sont les personnes en dessous de 40 ans qui expliquent le plus les arrêts, tandis qu'après 50 ans, les personnes ont moins d'arrêts.

Enfin, les performances obtenues pour ce modèle sont présentées :

➢ Performances globales

Gini	RMSE	DEVIANCE AVG	MAE
47,33%	2,22	1,54	0,71

Nous remarquons que les performances sont meilleures que le XGB qui présentait un Gini de 28%, un RMSE de 6,46, et un MAE de 1,58. L'amélioration du Gini est particulièrement importante. Ce modèle GLM est selectionné puis analysé plus en détail avec la courbe Lift.

➢ Courbe Lift

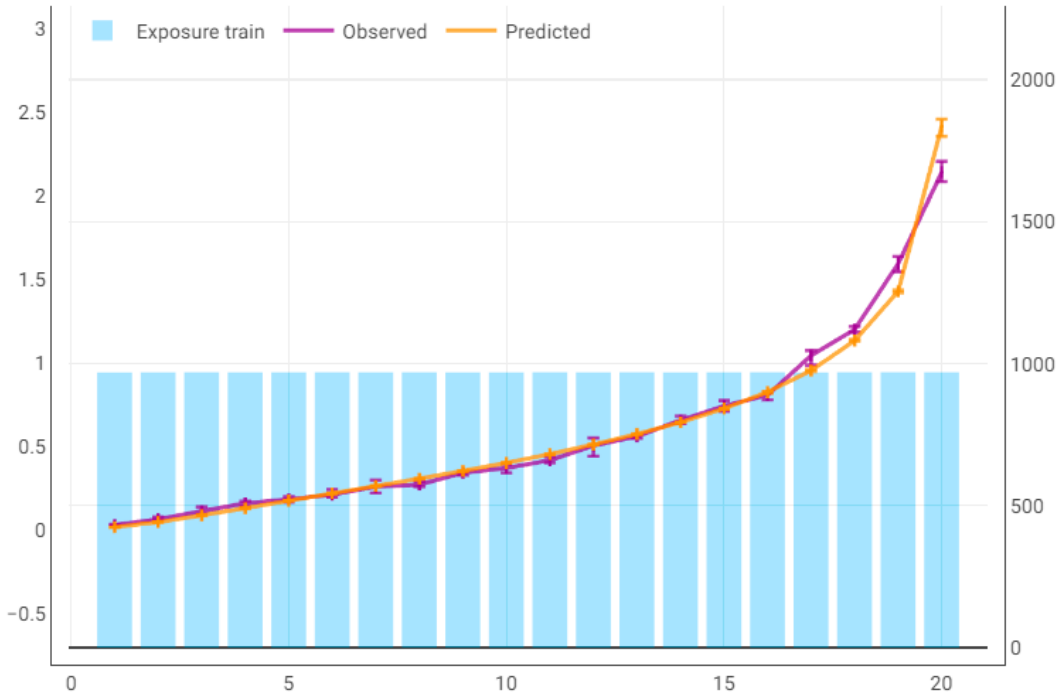


Figure 77 : courbe Lift pour le modèle de fréquence GLM

La courbe Lift nous montre que les prédictions sont moins bonnes pour les nombres d'arrêts importants car du 16ième au 19ième groupe, les observations sont légèrement au-dessus des prédictions et c'est l'inverse dans le dernier groupe.

Maintenant que les performances du modèle ont été analysées, les résidus quantiles normalisés randomisés[32] sont présentés. C'est une manière de vérifier la validité du modèle car dans un modèle GLM valide et dans une limite de volume de données, les résidus sont censés suivre une loi normale. Donc, les résidus quantiles normalisés randomisés doivent suivre en théorie une loi normale centrée réduite. Voici leur expression :

$$R = F^{-1}_{N(0,1)}(\hat{F}_i(Y_i))$$

Où $\hat{F}_i$ est la fonction de répartition estimée pour la réponse Y_i. Elle est randomisée ici car nous sommes dans le cas d'une loi de Poisson qui est discrète. En effet, si Y suit une loi dont la fonction de répartition est continue, alors $F_Y(Y)$ est une variable aléatoire qui suit la loi uniforme sur [0 ;1], et $F^{-1}_{N(0,1)}(F_Y(Y))$ suit donc la loi

[32] Peter K. Dunn and Gordon K. Smyth, Randomized quantile residuals, Journal of Computational and Graphical Statistics 5 (1996), no. 3, 236–244.

normale $N(0,1)$. Si Y est discrète, il n'est plus vrai que $F_Y(Y)$ suit la loi uniforme sur [0 ;1], mais elle peut être randomisée en faisant des tirages uniformes dans les sauts de la fonction de répartition. $F_Y(Y)$ est alors remplacée par une variable aléatoire Z telle que $Y = F_Y^{-1}(Z)$, où Z suit la loi uniforme sur [0 ;1].

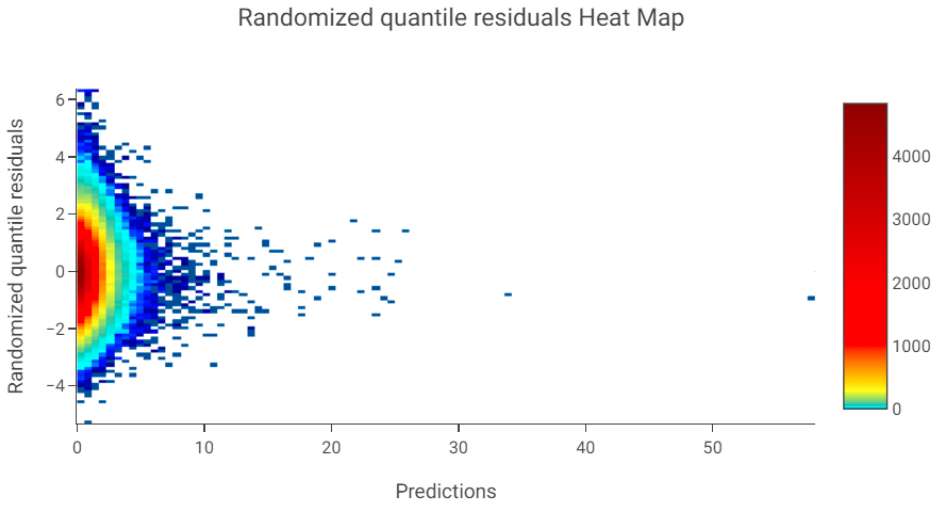


Figure 78 : quantiles normalisés randomisés pour la loi de fréquence

Nous observons que les résidus sont centrés autour de 0 et sont symétriques par rapport à l'axe horizontale d'ordonnée 0. L'échelle de couleurs représentant l'exposition, nous constatons que l'essentiel des résidus se situe dans la zone rouge dans l'intervalle [-2 ;2]. Les poids dans les régions bleues sont 1/5000 moins importants que dans les régions rouges. Nous pouvons donc valider le modèle qui se base bien sur l'hypothèse de la loi normale centrée réduite pour ces résidus.

Nous en venons à présent au modèle de durée et le choix du modèle n'est pas détaillé car il suit le même principe que pour le modèle de fréquence. Les résultats du modèle sont directement présentés.

Section 6.2 : modèle de durée

Le modèle de durée est construit en faisant l'hypothèse que Y suit une loi Gamma.

Pour le modèle de durée, voici la sélection des 12 variables obtenue:

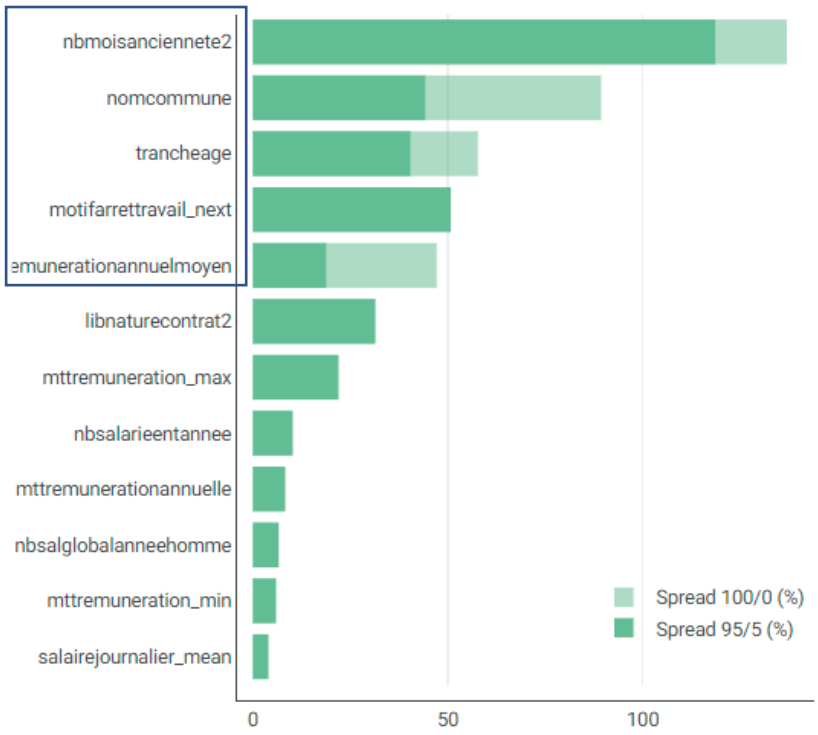


Figure 79 : classification des variables selon leur importance pour le modèle de durée

Nous regardons maintenant l'effet de quelques variables sélectionnées sur la durée de l'arrêt.

Tout d'abord, nous nous intéressons à l'effet de l'ancienneté du salarié dans l'entreprise sur la durée de l'arrêt.

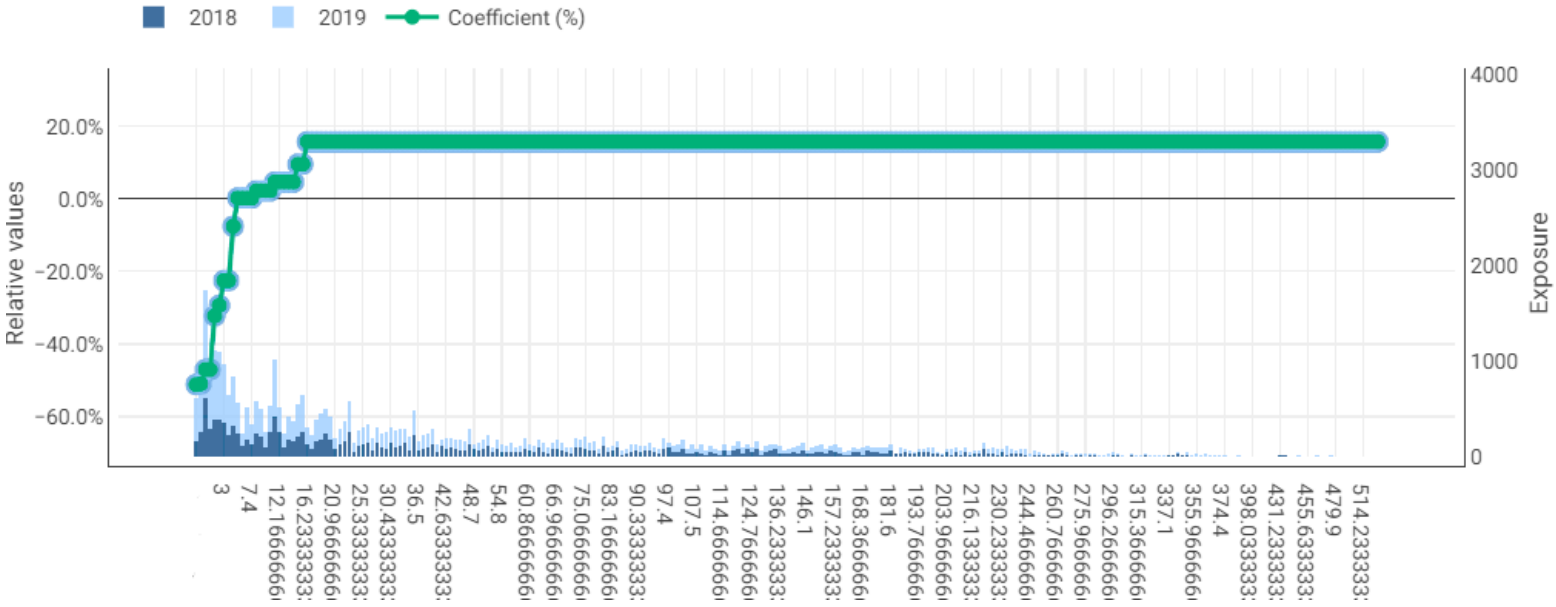


Figure 80 : effet de l'ancienneté du salarié dans l'entreprise (en mois) sur la durée d'arrêt

Nous constatons que les salariés qui ont une ancienneté inférieure à 7 mois ont des arrêts courts par rapport à la moyenne. Ce sont plutôt les salariés qui ont une ancienneté supérieure à 16 mois qui expliquent les arrêts longs. Cela s'explique par le fait qu'un salarié nouveau dans l'entreprise a moins de chance de faire de longs arrêts par rapport à un salarié qui est plus ancien.

Ensuite, l'effet de l'âge des salariés sur la durée d'arrêt est analysé.

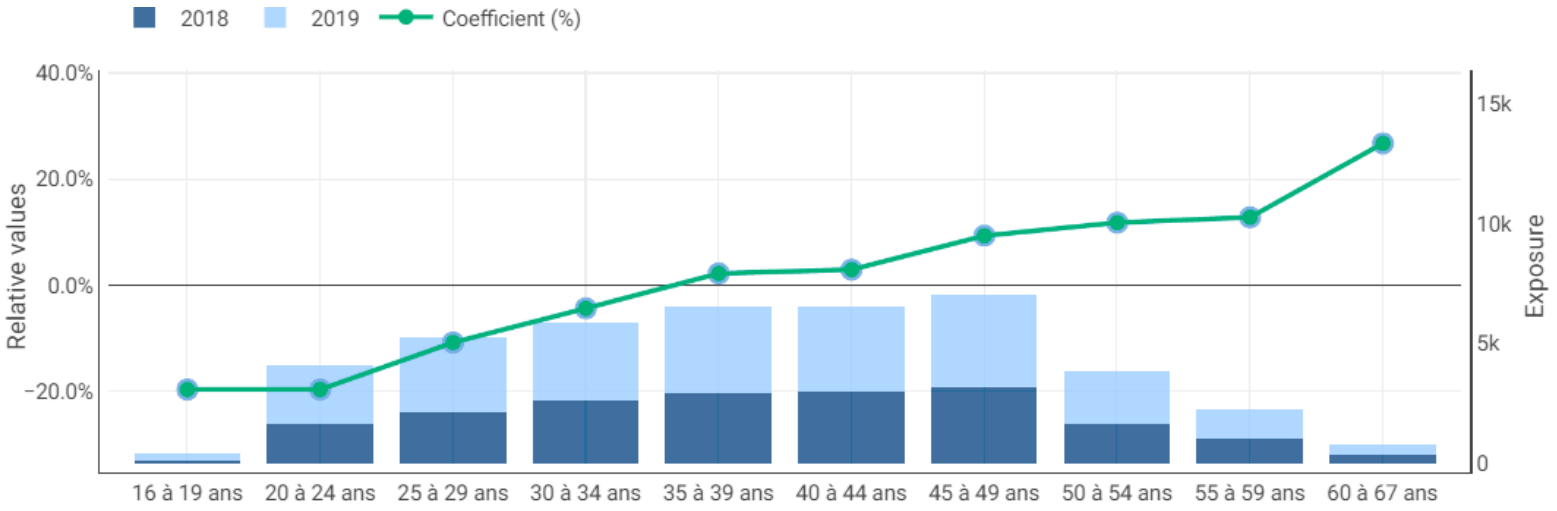


Figure 81 : effet des tranches d'âge sur la durée d'arrêt

Nous pouvons voir ici que les salariés qui ont le plus d'influence sur la durée des arrêts sont les salariés de plus de 35 ans. A partir de 35 ans, les durées sont globalement plus longues que pour les plus jeunes, et particulièrement longues après 45 ans.

Enfin, nous pouvons analyser l'effet du montant de rémunération annuel moyen par entreprise sur la durée d'arrêt.

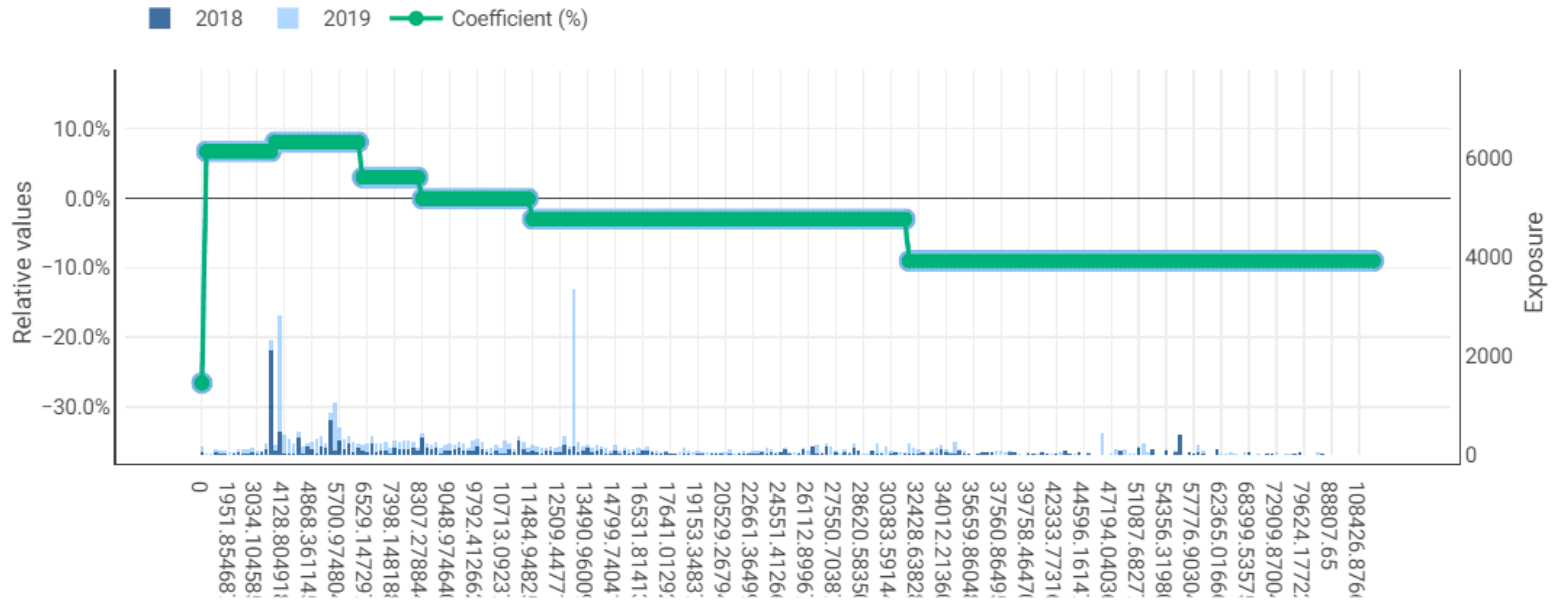


Figure 82 : effet du montant de rémunération annuel moyen par entreprise sur la durée d'arrêt

Nous observons que les montants de rémunération annuels moyens en dessous de 8 000 € environ et ceux de plus de 32 000 € environ ont plus d'effet sur la durée de l'arrêt que les autres montants. Cela s'explique par le fait que cette variable est construite sur la taille de l'entreprise. Comme vu dans les statistiques, les petites entreprises (moins de 50 salariés) ont des durées d'arrêts plus longues que les grandes.

Mais encore, nous pouvons regarder la cartographie des durées d'arrêts qui représente les prédictions des durées d'arrêts en fonction du lieu. Nous constatons que les durées sont particulièrement importantes dans les régions Rhône-Alpes et Provence-Côte d'Azur, en Corse, dans une partie de la Bretagne et au Nord-Pas-de-Calais. Les régions où les durées des arrêts sont les plus basses sont l'Alsace et le Centre.

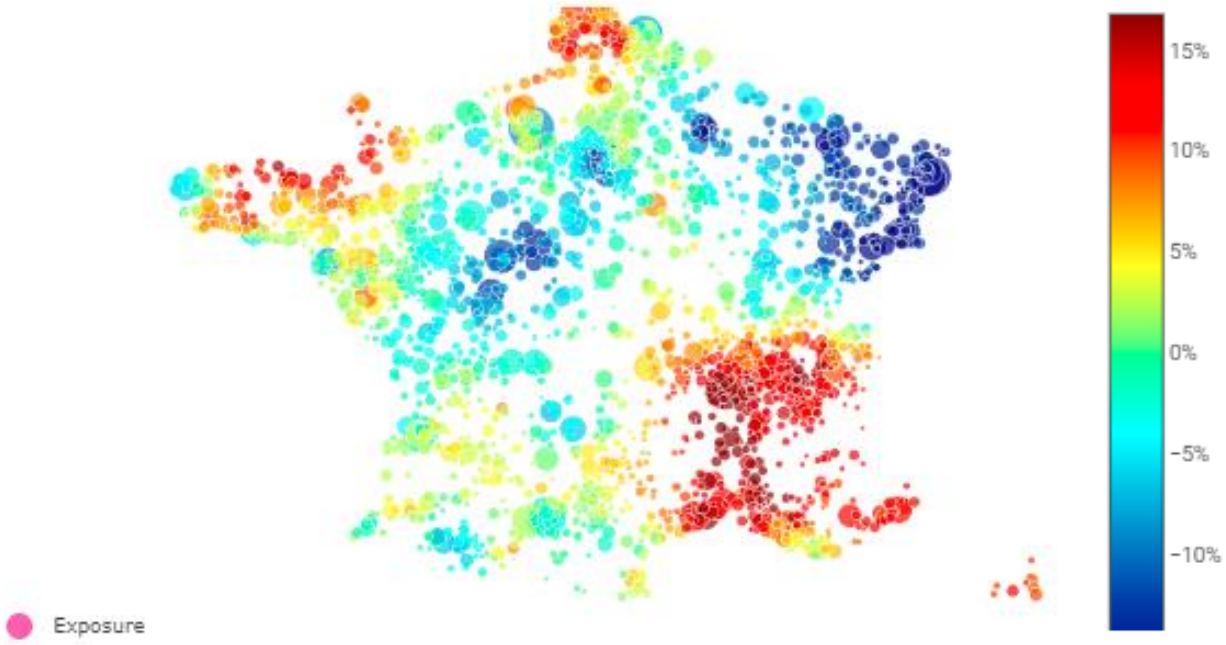


Figure 83 : cartographie des durées d'arrêts

Et enfin, les performances obtenues sont présentées.

- Performances globales

Gini	RMSE	DEVIANCE AVG	MAE
23,88%	23,61	1,28	13,6

Nous constatons que les performances sont meilleures que celles du XGB qui présentait un Gini de 11%, un RMSE de 23,70, et un MAE de 13,80. L'amélioration est encore une fois plus marquante au niveau du Gini.

- Courbe Lift

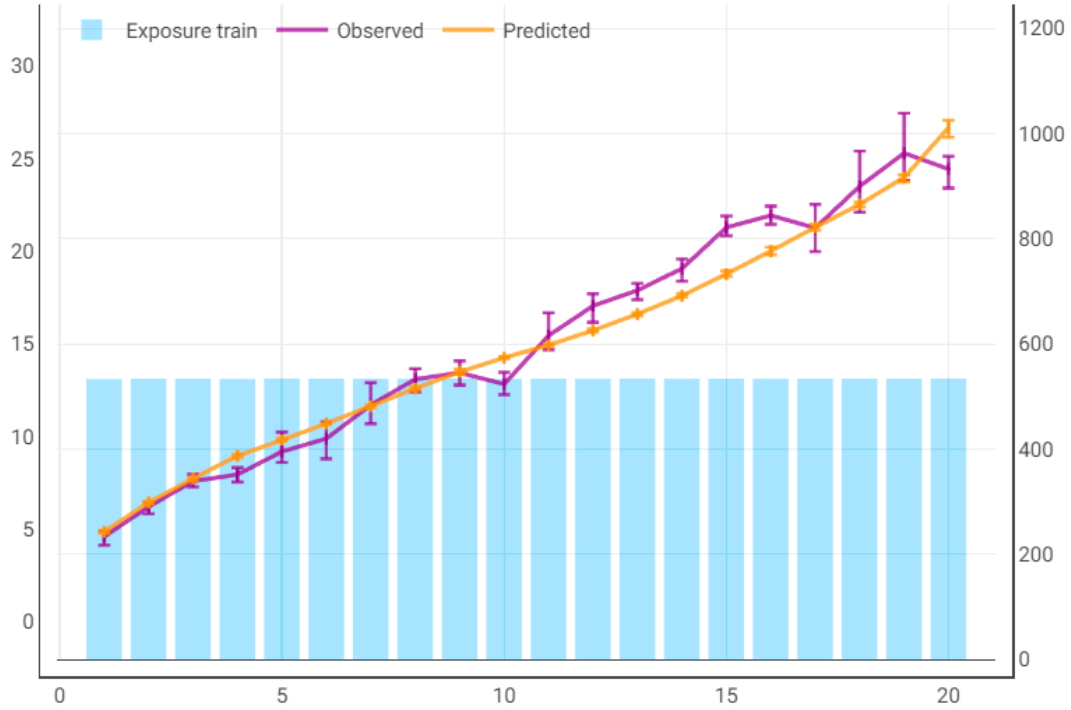


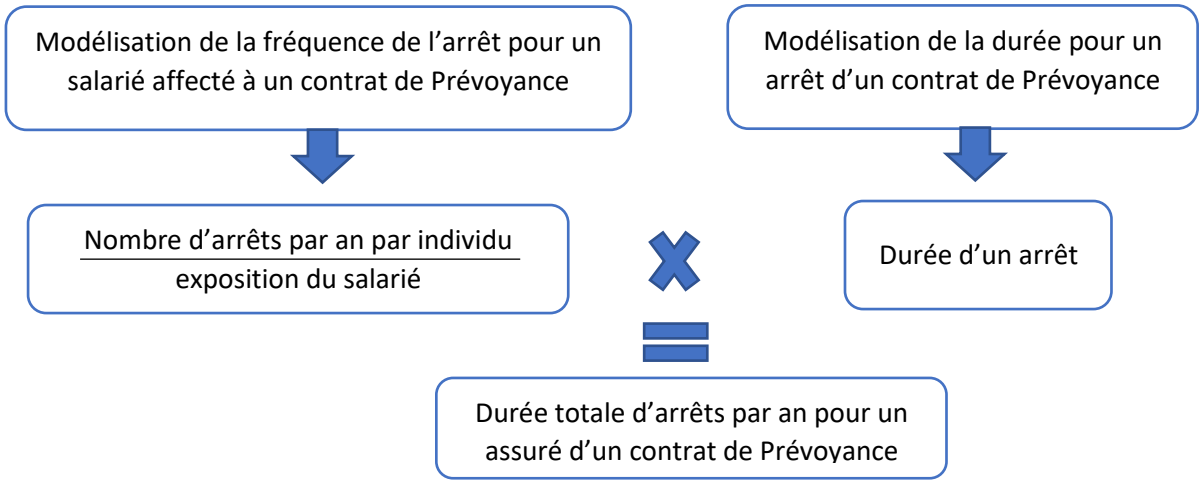
Figure 84 : courbe Lift pour le modèle de durée GLM

La courbe Lift prédit bien avant le 10ième groupe de durées. Cependant, à partir du 10ième groupe dont la durée observée moyenne correspond à 13 jours, le modèle prédit moins bien. Les prédictions sur les arrêts de plus de 13 jours (moyenné) manquent de précision à 2 ou 3 jours près.

Une fois que les lois de fréquence et de durée provenant des modèles GLM ont été choisies, les deux modèles peuvent être agrégés afin d'obtenir le modèle final qui correspond à la multiplication de ceux-ci.

Section 6.3 : modèle agrégé fréquence × durée

Le modèle agrégé correspond à la multiplication des prédictions de la fréquence et de la durée qui est schématisée ci-dessous.



Grâce au modèle agrégé, la durée totale des arrêts sur une période d'un an peut être prédite. Le coût de l'arrêt de travail peut donc être estimé et il est possible d'en sortir une tarification.

Partie 7 : tarification

Pour établir une tarification à partir du modèle agrégé, nous choisissons les variables qui ressortent comme les plus importantes dans nos modèles GLM de fréquence et de durée :

- APE
- Montant de rémunération annuel du salarié
- Effectif de l'entreprise
- Ancienneté du salarié
- Commune où se situe l'entreprise
- Age du salarié
- Rémunération totale de l'entreprise

Le tarif doit se faire au niveau collectif et non au niveau individuel. Les critères (variables) numériques sont donc agrégées à la moyenne sur l'entreprise : âge moyen, salaire moyen, etc.

Dans le tarificateur, il y a trois parties :

- Les listes de tous les coefficients correcteurs pour toutes les variables de tarification et leurs modalités ;
- Les données démographiques à remplir par le commercial ;
- Le niveau de garantie à remplir par le commercial ;
- Les hypothèses de calcul ;
- Le tarif obtenu.

Le tarificateur et son fonctionnement sous-jacent sont présentés.

➢ **<u>Les coefficients correcteurs</u>**

Le tarif est ajusté grâce à des coefficients correcteurs résultants du modèle GLM. Ils correspondent aux exponentielles de β, estimés par la méthode du maximum de vraisemblance et provenant de la formule du GLM ci-dessous :

$$\hat{Y} = \mathbb{E}\,[Y \mid X = \mathrm{x}] = g^{-1}\left(\sum_{\substack{i\epsilon[\![1,p]\!] \\ j\in[\![1,q_i]\!]}} x_{i,j} * \hat{\beta}_{i,j}\right) = \exp\left(\sum_{\substack{i\epsilon[\![1,p]\!] \\ j\in[\![1,q_i]\!]}} x_{i,j} * \hat{\beta}_{i,j}\right)$$

- $\hat{Y}$: la prédiction du modèle ;
- X : le vecteur des variables $X_{i\epsilon[\![1,p]\!]}$ et x son observation ;
- g : la fonction de lien log ;
- p : le nombre de variables ;
- q_i : le nombre de modalités associé à la variable X_i ;
- $x_{i,j}$: variable n°i pour sa modalité n°j encodée en binaire : 1 si X contient cette réalisation, 0 sinon ;
- $\hat{\beta}_{i,j}$: le coefficient estimée pour la modalité n°j de la variable X_i.

Pour mieux comprendre où interviennent les coefficients, la formule est illustrée avec des exemples de variables :

$$\text{Prédiction} = e^{intercept} \times e^{\beta_{l,m}} \times \ldots \times e^{\beta_{n,o}} \times \ldots$$

$$\text{Prédiction} = Constante \times Coefficient_{âge} \times \ldots \times Coefficient_{salaire} \times \ldots$$

Où $Constante$ est l'offset du modèle qui correspond au nombre de jours total « moyen » d'arrêts sur un an sans prendre en compte les effets des différentes variables.

- **Les données démographiques à saisir**

	A remplir	Coefficients correcteurs
APE	Fonderie de fonte	1,00
Montant de rémunération annuel moyen	20 000,00 €	1,30
Effectif	400	1,61
Ancienneté moyenne (mois)	3	0,75
Commune	ABBEVILLE	1,05
Age moyen	47	1,04
Rémunération totale entreprise	10 000 000,00 €	0,92

Correctifs en fonction des critères remplis

- **Niveau de garantie à saisir**

Le niveau de garantie est à renseigner par le commercial. Il représente le pourcentage du salaire annuel garanti.

Niveau de garantie	80%

- **Hypothèses**

Afin de simplifier le calcul de la tarification, plusieurs hypothèses sont émises sur les taux suivants :

Taux remboursement SS	50%
Taux remboursement employeur	0%
Chargements	10%

Les trois jours de carence précédant le remboursement de la Sécurité Sociale sont omis. Puis, le remboursement de l'employeur par la loi de mensualisation est considéré comme nul.

➢ **Calcul du tarif**

Probabilité d'arrêt ajustée	5,90%
Rémunération moyenne	25 000,00 €
Niveau de garantie	80%

Coût total	Remboursement SS	Remboursement employeur	Prime Pure %
4,72%	2,95%	0%	1,77%

Prime Pure €
442,63 €
Prime commerciale
491,82 €

Afin d'aboutir à la valeur de la prime pure, nous procédons en plusieurs étapes de calcul.

1- Probabilité d'arrêt = Offset / 365 où Offset = nombre de jours total « moyen » d'arrêts sur un an

2- Probabilité d'arrêt ajustée = Probabilité d'arrêt × $Coefficient_{APE}$ × $Coefficient_{MONTANT\ REMUNERATION\ ANNUEL\ MOYEN}$ × $Coefficient_{EFFECTIF}$ × $Coefficient_{ANCIENNETE\ MOYENNE}$ × $Coefficient_{COMMUNE}$ × $Coefficient_{AGE\ MOYEN}$ × $Coefficient_{REMUNERATION\ TOTALE\ ENTREPRISE}$

3- Rémunération moyenne (annuelle) = Rémunération totale de l'entreprise / Effectif de l'entreprise

4- Coût total (en pourcentage du salaire) = Niveau de garantie × Probabilité d'arrêt ajustée

5- Prime pure (%) = Coût total – Remboursement SS – Remboursement employeur

6- Prime pure (€) = Prime pure (%) × Rémunération moyenne annuelle

7- Prime commerciale = $\frac{\text{Prime pure(€)}}{1-\text{chargements}}$

L'objectif était de challenger le tarificateur déjà présent au sein de l'entreprise. L'outil créé dans ce mémoire mais qui doit encore être amélioré présente les évolutions suivantes :

- Les coefficients correcteurs ajustant les tarifs sont plus précis car ils sont calculés sur une base qui contient des sinistrés mais aussi des non-sinistrés ainsi que des arrêts post-franchise, ce que les tables réglementaires ne possédaient pas.
- Les tarifications sont à présent établies d'après des données de notre portefeuille.
- Les variables de tarification ne sont pas exactement les mêmes qu'avant (commune, ancienneté non présentes…). Et, les effets d'interaction entre variables (variables croisées, exemple : sexe × âge) n'apparaissent pas comme déterminants pour la tarification alors qu'ils étaient utilisés dans l'ancien outil.
- L'outil est facilement actualisable grâce aux modélisations qui peuvent être mises à jour sur des données qui peuvent changer.
- Les modèles performants de Machine Learning ont permis d'obtenir des prédictions plus précises et donc une tarification plus fiable qu'auparavant.

Conclusion

Dans le portefeuille de Prévoyance collective étudié, les données du suivi du risque de l'arrêt de travail sont issues des systèmes de gestion internes. Pour un portefeuille sans affiliation, cela correspond donc aux données de salariés sinistrés, avec des sinistres au-delà de la franchise contractuelle.
Or, depuis 2017, les flux de données provenant de la DSN sont reçus par les organismes complémentaires : ceci leur donne accès entre autres à tous les évènements marquant la vie du salarié rattaché au contrat de Prévoyance, en particulier, tous les arrêts de travail auxquels il fait face (même pour les durées avant franchise). Grâce à cela, les actuaires des assureurs accèdent à une vision plus complète pour étudier le risque d'absentéisme.
Ceci est d'autant vrai pour AG2R, une IP qui assure des branches sur le risque « maintien de salaire » : il s'agit d'arrêts courts non modélisés par les tables réglementaires BCAC.

Rappelons que l'objectif du mémoire est de challenger les normes tarifaires des produits de la Prévoyance collective concernant les arrêts de travail grâce aux données de la DSN et grâce aux modèles de Machine Learning.

Le risque d'arrêt a d'abord été analysé. Les données de la DSN 2018 et 2019 relevant des branches « aide à domicile » et « propreté » ont été choisies car elles représentent un enjeu particulier par leur volume d'arrêts et par leur population particulière.

Ensuite, la loi de maintien du BCAC a été comparée avec l'estimation de la loi sur nos données par la méthode de Kaplan Meier. Par une simulation sur la durée de franchise, il a été constaté qu'intégrer les arrêts dont la durée ne dépasse pas la franchise permet de modéliser une loi différente et plus précise. L'apport de la DSN dans notre étude a ainsi été démontré.

Cependant, cette méthode est basée seulement sur le paramètre de l'âge et nous tenions à prendre en compte autant de caractéristiques du salarié et de ses arrêts que possible. Plusieurs modélisations des lois de fréquence et de maintien en arrêt de travail ont donc été testées à l'aide des méthodes de Machine Learning : les algorithmes CART, Random Forest et XGBoost. D'après les indicateurs de RMSE, MAE, Gini, les courbes de Lorenz et les courbes Lift, le modèle du XGBoost a été sélectionné comme le plus performant de ces modèles.

Par ailleurs, lors de cette étude, nous avons abordé un enjeu important pour l'IP : comment maintenir des normes tarifaires à jour et comment passer d'un modèle construit manuellement sur Python à un modèle prêt à l'emploi rapidement. Une évaluation d'un outil de tarification externe a été donc menée. Il a été constaté que ses performances dépassent celles du XGBoost retenu sur le premier volet de l'étude. C'est pourquoi, les lois de fréquence et de maintien construites avec le GLM ont finalement été adoptées plutôt que celles du XGBoost. Pour les valeurs élevées, les prédictions restent imprécises avec une erreur allant jusqu'à 0,5 arrêt (valeur moyenne) pour les fréquences extrêmes, et une erreur allant jusqu'à 3 jours d'arrêt (valeur moyenne) pour les durées importantes.

Une des limites de notre étude se trouve alors dans la modélisation de la durée, qui pourrait encore être améliorée. Pour améliorer sa précision de prédiction, il serait envisageable de transformer la variable numérique de la durée de l'arrêt en une variable catégorielle avec des tranches de durée, la perte de précision provenant probablement de la forte dispersion des valeurs de la durée. Il est aussi possible de procéder non plus à des algorithmes de régression mais à de la classification.

Au regard des résultats des GLM, nous avons retenu les variables les plus explicatives, pour les modèles de fréquence et de durée, comme critères de tarification : l'ancienneté, l'âge, le lieu de travail (commune), le secteur d'activité (APE), le montant de rémunération annuel du salarié, l'effectif de l'entreprise et la rémunération totale de l'entreprise. Les coefficients du modèle GLM associés aux modalités de ces variables et obtenus par maximisation de la vraisemblance nous ont servi à ajuster nos tarifs au profil du client.

La tarification a été conçue à partir du modèle qui agrège les deux lois : « fréquence d'arrêt sur un an » × « durée d'un arrêt ». En fonction d'un profil client entreprise, une probabilité d'arrêt a été calculée en multipliant une probabilité d'arrêt moyenne multipliée elle-même par des coefficients correcteurs. En multipliant encore cela par le niveau de garantie et le montant de rémunération annuel moyen, le tarif a été obtenu.

La tarification a donc été challengée en testant plusieurs modèles de Machine Learning et en poussant la performance au mieux. Ces modèles ont jugé certaines variables importantes pour la tarification et qui n'apparaissaient pas forcément dans le tarificateur précédent. Mais aussi, la précision de la tarification a pu être améliorée en créant nos lois sur des données plus fines au niveau du salarié, mais surtout en élargissant le périmètre à celui des non-sinistrés et des arrêts dont la durée est inférieure à celle de la franchise. Enfin, il est à noter que ce nouveau tarificateur est facilement actualisable grâce aux modélisations qui peuvent s'adapter au portefeuille, ce qui n'était pas le cas auparavant avec un tarificateur moins flexible.

D'autres pistes sont intéressantes pour approfondir notre étude.
En premier lieu, nous constatons que notre tarificateur ne prend pas en compte certaines variables qui sont d'habitude utilisées comme critères de tarification comme le sexe et la CSP. Il serait pertinent de forcer ces variables dans les modélisations pour faire face au besoin métier et pour analyser l'impact de ces variables.
En second lieu, pour démontrer l'avantage de l'utilisation du volume particulièrement important des données de la DSN, nos modélisations pourraient être effectuées sur les arrêts longs de plus de 6 mois. Il serait ainsi possible de comparer nos lois à celles construites sur les bases ne provenant pas de la DSN, cette fois-ci sur les arrêts longs. Ce dernier point ne peut pas être réalisé actuellement car il n'est pas possible d'utiliser les années suivant 2019 touchées par le Covid. Elles viendraient biaiser nos modélisations.

Bibliographie

Mémoires

ALI A. (2016) : « Invalidité et arrêt de travail en Australie »

CAUX J. (2017) : « Modélisation de l'entrée en incapacité temporaire de travail en Prévoyance collective »

JOUANIGOT E. (2009) : « Etude comparative de modèles de durée en présence de données hétérogènes et censurées »

JOUITTEAU A. (2013) : « Construction de lois de maintien en arrêt de travail en assurance emprunteur »

KARAMOKO FOFANA K. (2017) : « Approche tarifaire des contrats collectifs Frais de Santé à l'aide des méthodes d'apprentissage »

KOY G. (2019) : « Comparaison des méthodes classiques et alternatives avec le Machine Learning pour la construction d'une table de mortalité d'expérience Best Estimate »

LAGOS T. (2018) : « Étude de méthodes innovantes de Machine Learning permettant la tarification de produits Santé en mobilité internationale «

LEBOSSE C. (2009) : « Construction de barèmes de tarification d'arrêt de travail par une méthode de Simulation – Application à un portefeuille de Travailleurs Non-Salariés »

LECOMPTE K. (2018) : « Automatiser la comparaison de modèles : Application sur l'amélioration d'un modèle de fréquence par des techniques de Machine Learning »

RAHMOUNI A. (2016) : « Construction de lois expérimentales en incapacité temporaire de travail »

ROGER Y. (2017) : « Refonte des lois biométriques dans une approche « Best Estimate » de la modélisation de l'arrêt de travail »

STEFANI L. (2019) : **«** Méthodes de construction des barèmes tarifaires collectifs en arrêt de travail et apport des données de DSN »

Ouvrages et documents

ALEANDRI M. (2019) : « Data Science in Insurance, Some introductory case studies »

AZENCOTT C. (2018) : « Introduction au Machine Learning »

GOLDBURGM., FELDBLUM S., KHARE A., and TEVET D. (2016) : « Generalized linear models for insurance rating, Casualty Actuarial Society »

LOPEZ O., MILHAUD X., THEROND P. (2016) : « Tree-based censored regression with applications in Insurance »

NIELSEN D. (2016) : « Tree Boosting With XGBoost : Why Does XGBoost Win "Every" Machine Learning Competition ? »

PING WANG, YAN LI, CHANDAN K. REDD (2017) : « Machine Learning for Survival Analysis : A Survey »

PLANCHET F. (2021) : « Modèles de durée »

Sites internet

Arbres binaires de décision : https ://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-cart.pdf

Site de BERARD J. : modèles linéaires généralisés : http ://irma.math.unistra.fr/~jberard/GLM-transp.pdf

Site de RAKOTOMATALA R. : http ://eric.univ-lyon2.fr/~ricco/cours/slides/gradient_boosting.pdf ,

https ://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_ACP_Python.pdf

Site de l'Argus de l'assurance : www.argusdelassurance.com

Site de la déclaration sociale nominative : www.dsn-info.fr

Site de l'INSEE : www.insee.fr

Site de Légifrance : www.legifrance.gouv.fr

Site de PLANCHET F. : www.ressources-actuarielles.net

Explication de la loi Evin : https ://pro.apicil.com/actualites/loi-evin-prevoyance-definition/

Cahier technique DSN 2020 : https ://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2020.pdf

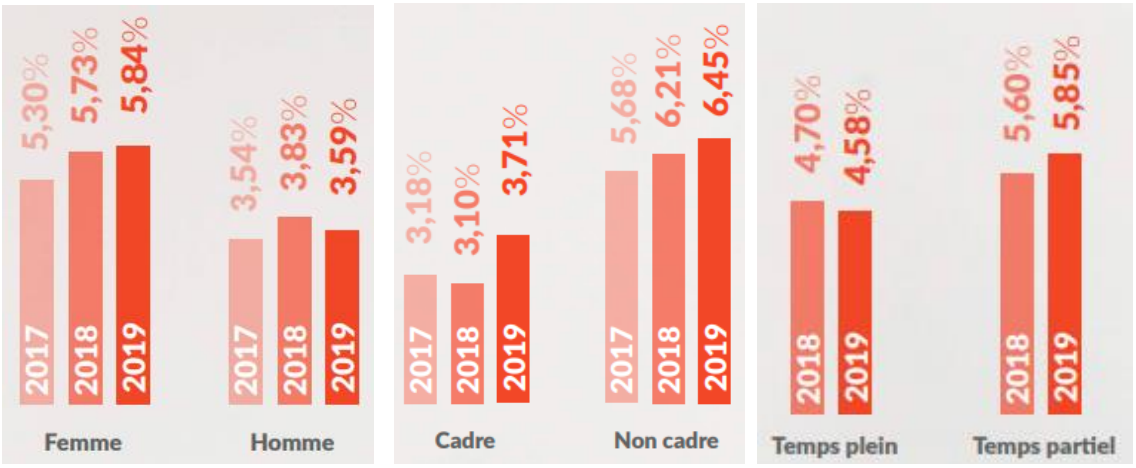
Table des figures

Table des annexes

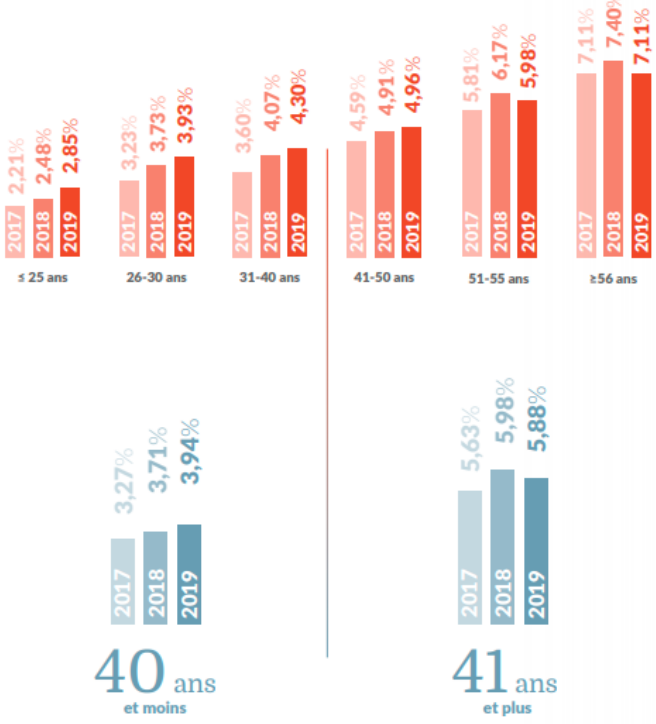
Annexes

Annexe 1 : taux d'absentéisme par sexe, statut, temps de travail et âge provenant du Baromètre Ayming 2020

Taux d'absentéisme en fonction du sexe, du statut et du temps de travail



Taux d'absentéisme en fonction de l'âge



Annexe 2 : article 1 de la loi Evin

Modifié par Ordonnance n°2018-470 du 12 juin 2018 - art. 5

« Les dispositions du présent titre s'appliquent aux opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité ou des risques d'incapacité de travail ou d'invalidité ou du risque chômage.

Seuls sont habilités à mettre en œuvre les opérations de couverture visées au premier alinéa les organismes suivants :

a) Entreprises régies par le code des assurances ;

b) Institutions de prévoyance relevant du titre III du livre IX du code de la sécurité sociale ;

c) Institutions de prévoyance relevant de la section 4 du chapitre II du titre II du livre VII du code rural ;

d) Mutuelles relevant du code de la mutualité.

e) Organismes visés aux articles L. 644-1 et L. 652-1 du code de la sécurité sociale pour les opérations mises en place dans le cadre des dispositions de l'article L. 144-1 du code des assurances. »

Annexe 3 : tableau des variables de la DSN nécessaires à notre étude

Tables	Variables	Contenu de la variable
Entreprise S21.G00.06	Entreprise.Siren	Numéro de SIREN de l'entreprise
	Entreprise.Apen	Code APE de l'entreprise
	Entreprise.EffectifMoyen	Effectif moyen de l'entreprise présent au 31/12 de l'année concernée
	Entreprise.CodePays	Code de la région géographique de l'entreprise (DOM-TOM, Monaco …)
Etablissement S21.G00.11	Etablissement.Apet	Code APE de l'établissement concerné
	Etablissement.CodePostal	Code postal de l'établissement
	Etablissement.EffectifFinPeriode	Effectif moyen de l'établissement présent au 3112 de l'année concernée
	Etablissement.CodePays	Code de la région géographique de l'établissement (DOM-TOM, Monaco …)

Individu S21.G00.30	Individu.Sexe	Sexe de l'assuré
	individu.DateNaissance	Date de naissance de l'assuré
	Individu.CodePostal	Code postal du lieu de résidence de l'assuré
	Individu.CodePays	Code de la région géographique de l'assuré (DOM-TOM, Monaco …)
	Individu.StatutEtranger	Statut Frontalier / étranger
	Individu.NiveauFormationPlusEleveIndividu	Niveau d'étude de l'assuré
Pénibilité S21.G00.34	Penibilite.FacteurExposition	Facteur de pénibilité de l'emploi de l'assuré
Contrat (contrat de travail, convention, mandat) S21.G00.40	Contrat.DateDebut	Date de début de contrat de travail
	Contrat.StatutConventionnel	Statut contrat
	Contrat.StatutRC	Statut catégoriel retraite complémentaire
	Contrat.PcsEse	CSP
	Contrat.LibelleEmploi	Fonction du salarié
	Contrat.Nature	Type de contrat (CDD/CDI …)
	Contrat.DateFinPrevisionnelle	Date de fin prévisionnelle du contrat de travail
	Contrat.ModaliteTemps	Temps partiel/Temps plein
	Contrat.ComplementBase	Type de régime de base de l'assuré
	Contrat.Ccn	Nom de la CCN associée
	Contrat.RegimeMaladie	Type de régime spécial
	Contrat.Lieutravail	Code lieu de travail de l'assuré
	Contrat.RegimeVieillesse	Identifiant du régime de base vieillesse
	Contrat.MotifRecours	Motif de l'embauche
	Contrat.StatusEmploi	Catégorie d'emploi (fonctionnaire, …)
	Contrat.CodeRisqueAccidentTravail	Code du risque accident du travail pour lequel l'assuré doit obligatoirement être assuré
Rémunération S21.G00.51	Remuneration.Type	Salaire de base/Heures supplémentaires …
	Remuneration.NombreHeures	Nombre d'heures concernées
	Remuneration.Montant	Montant de la rémunération
	TravailArret.Motif	Motif de l'arrêt de travail

Arrêt de travail S21.G00.60	TravailArret.DernierJour	Dernier jour travaillé (+1 : date d'arrêt)
	TravailArret.DateFinPrevisionnelle	Date de fin prévisionnelle de l'arrêt
	TravailArret.RepriseDate	Date de reprise du travail
	TravailArret.RepriseMotif	Motif de la reprise (temps partiel, …)
Fin du contrat S21.G00.62	ContratFin.DateFin	Date de fin de contrat
	ContratFin.Motif	Motif de la rupture du contrat
Autre suspension de l'exécution du contrat S21.G00.65	ContratSuspensionAutre.Motif	Autre motif de suspension (Inval, …)
	ContratSuspensionAutre.DateDebut	Date de début de la suspension
	ContratSuspensionAutre.DateFin	Date de fin de la suspension
Lieu de travail ou établissement utilisateur S21.G00.85	TravailLieu.Apet	Code APE de l'établissement de travail
	TravailLieu.CodePostal	Code postal de l'établissement
	TravailLieu.Localite	Localité établissement
	TravailLieu.Nature	Etablissement/Domicile/Autre …
	TravailLieu.CodeINSEEcommune	Code INSEE de la commune
Cotisation individuelle S21.G00.81	CotisationIndividuelle.MontantRéductionExonération	Organisme complémentaire : Montant total de la cotisation pour le salarié, au titre de la période de rattachement et de l'affiliation Prévoyance renseignées dans le bloc S21.G00.78
S21.G00.82 Cotisation établissement	CotisationEtablissement.Valeur	Organisme complémentaire : montant de cotisation dont la nature est renseignée dans la rubrique Code de cotisation S21.G00.82.002, due au titre de l'établissement déclaré en S21.G00.11

Annexe 4 : lexique des variables

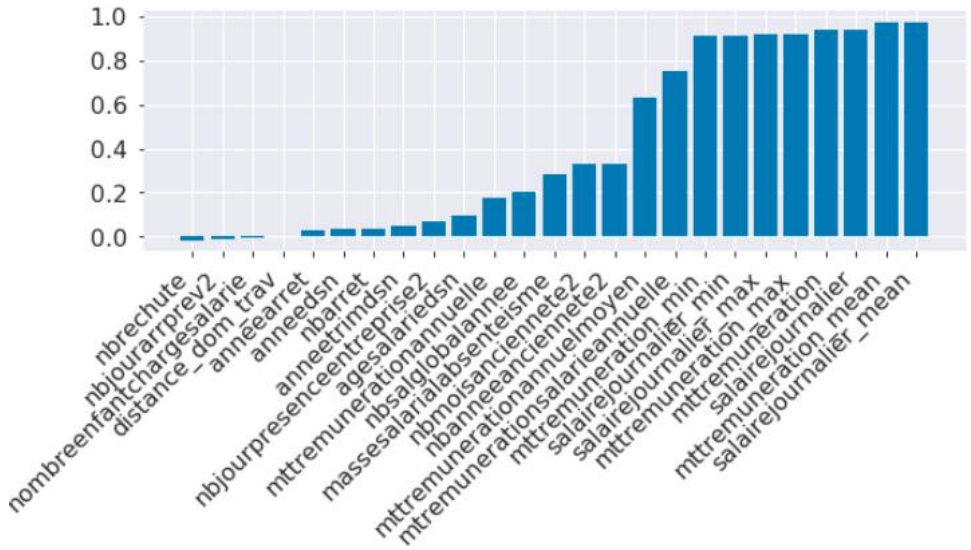
Variables	Explications	Règles de gestion
trancheetablissement	Tranches de taille d'établissement	Calcul du nombre de salariés par SIREN
anneedsn	L'année de déclaration DSN	X
anneetrimdsn	Dernier trimestre de déclaration disponible	X
Cddelegataireprevadh2	Code du délégataire de l'adhésion Prévoyance	X
moisdeclaration	Date de déclaration DSN	X
nomregionarret	Nom de la région de l'entreprise où a eu lieu l'arrêt	X
nomcommunearret	Nom de la commune où a eu lieu l'arrêt	X
dtnaissancesalaire	Date de naissance du salarié	Non renseignée
agesalariedsn	Age du salarié	Age du salarié à la déclaration DSN
trancheage	Tranches d'âge	Tranches d'âge des salariés à la déclaration DSN
sexe	Sexe du salarié (variable binaire)	X
lbsexe	Libellé homme ou femme du salarié	X
longituDegps	Longitude de l'entreprise de rattachement	X
latitudegps	Latitude de l'entreprise de rattachement	X
nomregion	Région de l'entreprise de rattachement	X
nomcommune	Commune de l'entreprise de rattachement	X
dtdebctrt_	Date de début du contrat	X
dtfinctrt_	Date de fin du contrat	X
mttremuneration	Montant de rémunération mensuel brut comprenant les primes	Calculé au moment de la déclaration
mttremunerationsalarieannuelle	Montant de rémunération annuel brut comprenant les primes	X
tranchesalaire	Tranches de salaire	Tranches effectuées sur « mttremuneration »
dtdernierjour	Date de dernier jour travaillé	X
dtfinprevisionnel	Date prévue de fin de l'arrêt de travail	X
dtreprise	Date de reprise effective	X
motifarrettravail	Motif de l'arrêt de travail	X
motifreprise	Motif de reprise du travail	X
trimestrearret	Trimestre de l'arrêt	X

Nbjourarrprev2	Nombre de jours d'arrêts	Différence entre la date de dernier jour travaillé « dtdernierjour » et la date de reprise du travail « dtreprise » pour chaque arrêt
Nbanneeanciennete2	Ancienneté dans l'entreprise en années	Différence d'années entre « moisdeclaration » et « dtdebctrt_»
Nbmoisanciennete2	Ancienneté dans l'entreprise en mois	Différence de mois entre « moisdeclaration » et « dtdebctrt_»
trancheanciennete	Tranche d'ancienneté	Calculée que pour les arrêts : Tranches faites sur « nbmoisanciennete »
anneearret	Année de l'arrêt	X
moisarret	Mois de l'arrêt	X
arret	Variable binaire indiquant 0 ou 1 si présence d'arrêt ou non	X
Libmotifarrettravail	Libellé du motif de l'arrêt de travail	X
Libmotifreprise_next	Libellé du motif de la reprise du travail	X
Libcdstatutrcsalarie2	Statut du salarié	X
Libstatutconventionnelsalarie2	Statut conventionnel du salarié	X
libmotiffinctrt	Libellé du motif de fin de contrat	X
libnaturecontrat	Libellé de la nature du contrat	X
nbsalglobalannee	Nombre de salariés par année de DSN et lieu de travail	X
nbsalglobalanneefemme	Nombre de femmes par année de DSN et lieu de travail	X
nbsalglobalanneehomme	Nombre d'hommes par année de DSN et lieu de travail	X
salairejournalier	Salaire journalier	« mttremuneration » divisé par 30
massesalarialabsenteisme	Masse salariale de l'absentéisme	« nbjourarrprev » × « mttremuneration » / 30
nbarret	Nombre d'arrêts par contrat et par année DSN	X
nbrechute	Nombre de rechutes par contrat et par année DSN	X
Nbjourpresenceentreprise2	Nombre de jours calendaires de présence en entreprise par salarié (pour un contrat) et par année	X
Distance_dom_trav	Distance du domicile de l'assuré à son lieu de travail	X
Mtremuneration_max/min/mean	Montant de rémunération mensuel brut maximum/minimum/moyen de	X

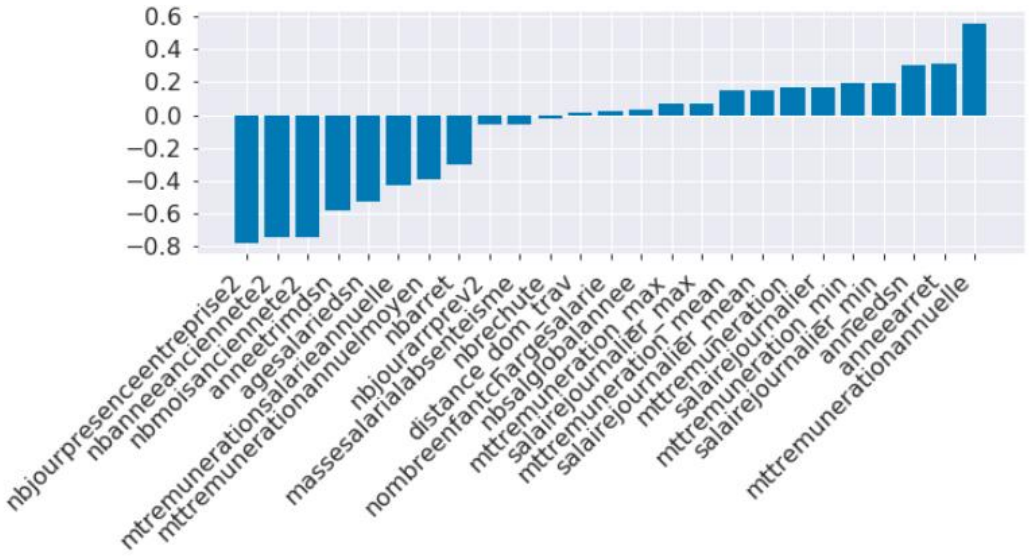
	l'assuré (pour un contrat) par année	
salairejournalier_max/min/mean	Salaire journalier brut maximum/minimum/moyen de l'assuré (pour un contrat) par année	X
mtremunerationannuelmoyen	Montant de rémunération annuel brut moyen par entreprise et par année	X
Nombreenfantchargesalarie	Nombre d'enfants à charge	X
tranchemtremunerationsalarieanuelle	Tranches pour les rémunérations annuelles brutes	X
embauche	Modalité de temps de travail dans l'entreprise	X
Regimealsacemoselle	Bénéficiaire du régime Alsace Moselle ou non	X
Secteur_etab	Secteur lourd, moyen ou léger de l'établissement	X
Secteur_ent	Secteur lourd, moyen ou léger de l'entreprise	X

Annexe 5 : corrélations des variables quantitatives avec les composantes principales

Corrélations avec le premier axe

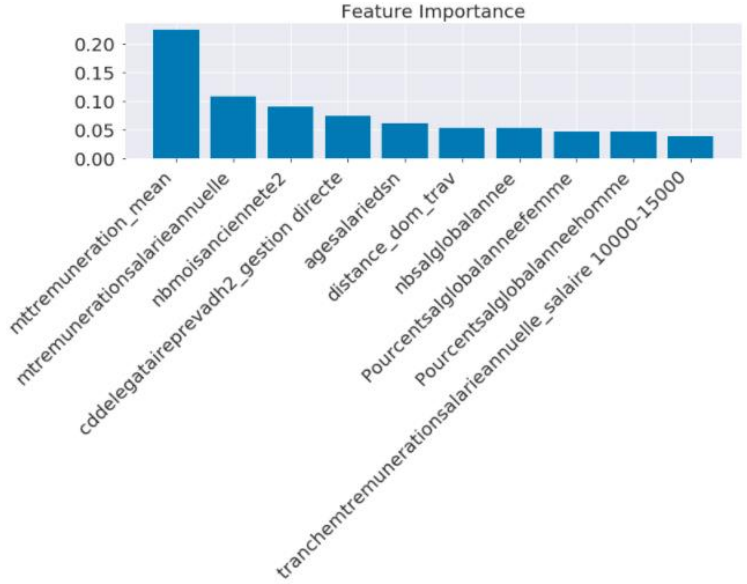


Corrélations avec le deuxième axe

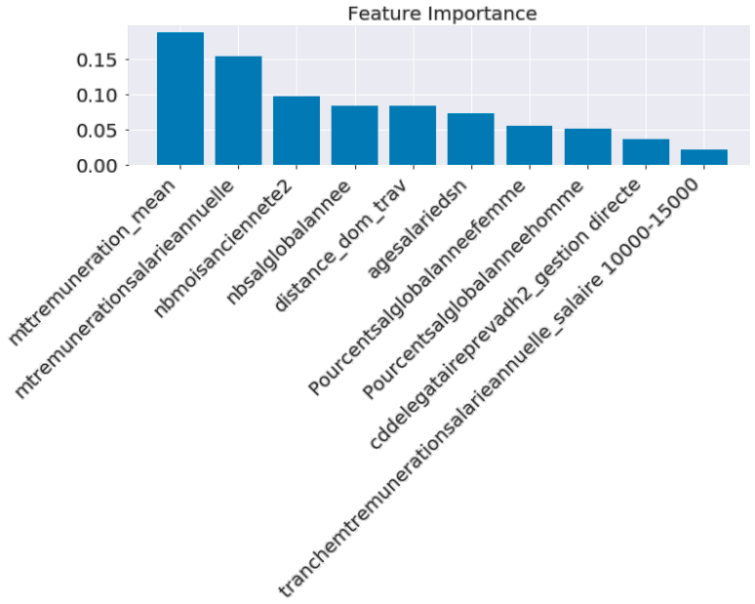


Annexe 6 : importances des variables pour les modèles de fréquence et de durées pour le CART et le Random Forest

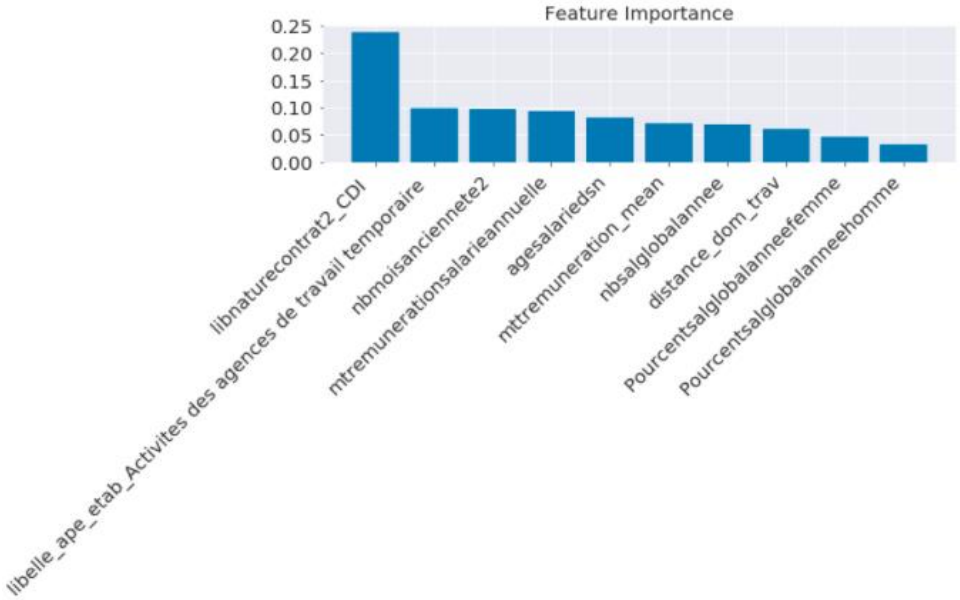
CART pour le modèle de fréquence



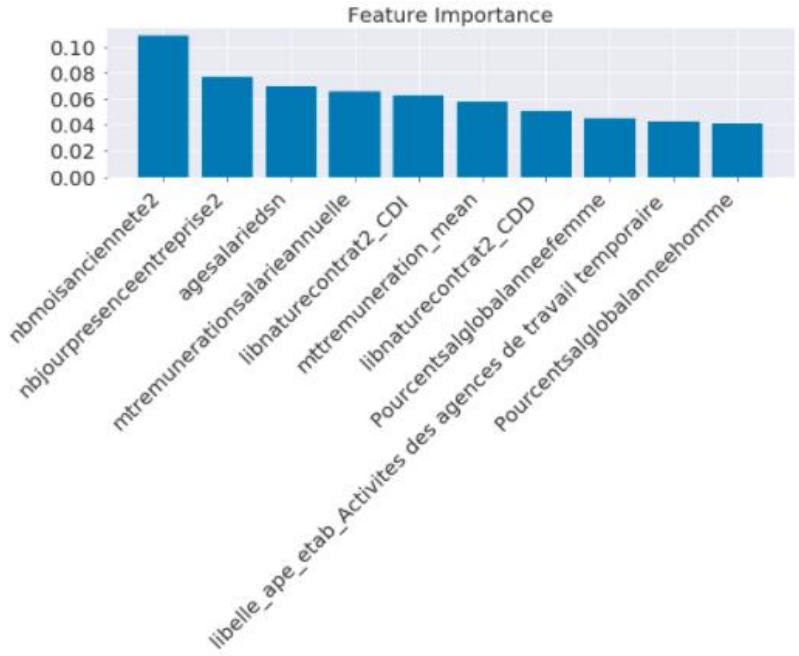
Random Forest pour le modèle de fréquence



CART pour le modèle de durée

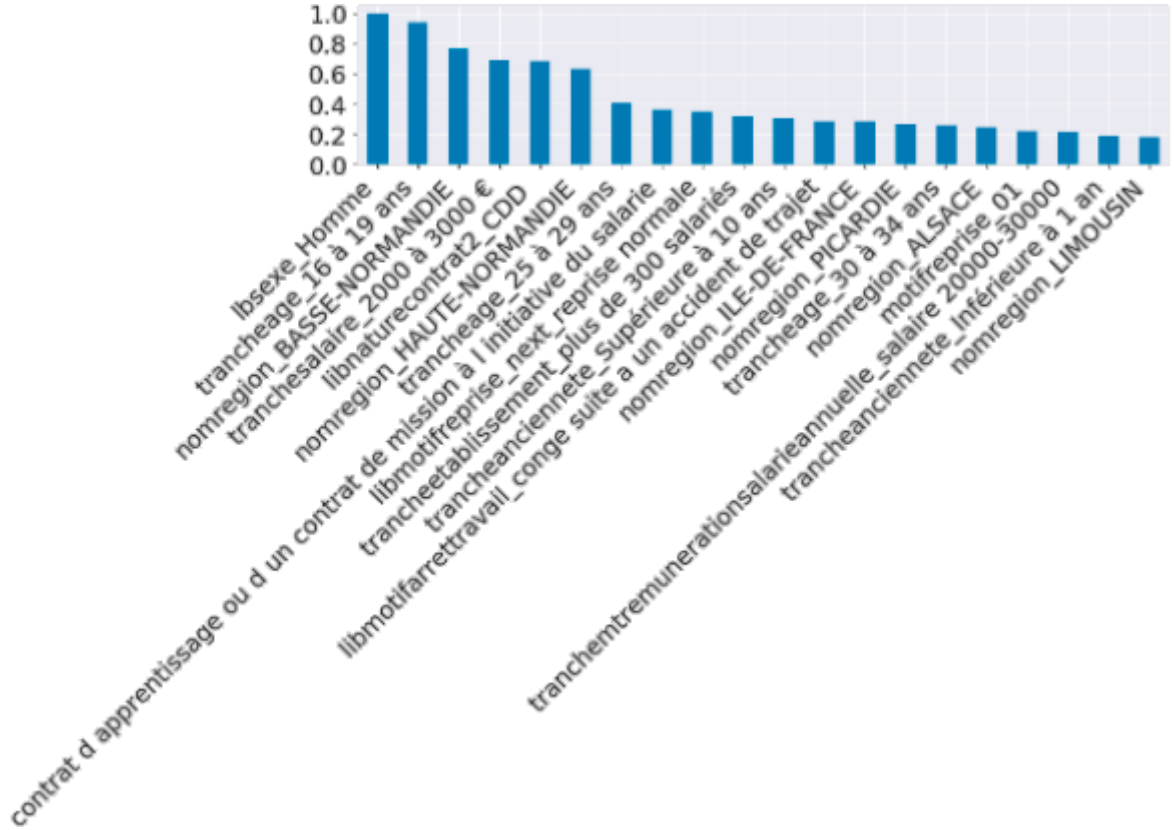


Random Forest pour le modèle de durée

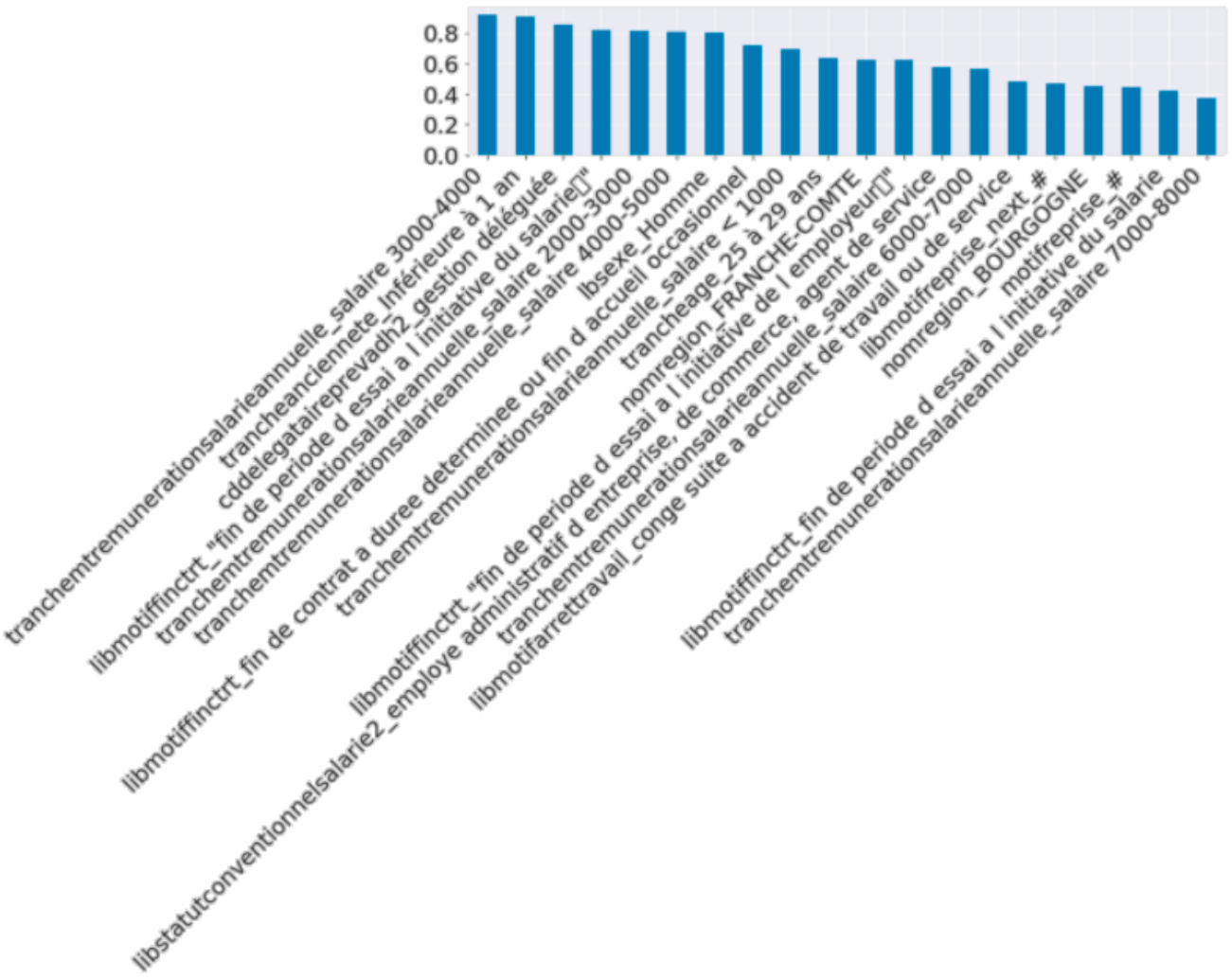


Annexe 7 : ACM - coordonnées des variables par rapport aux composantes principales

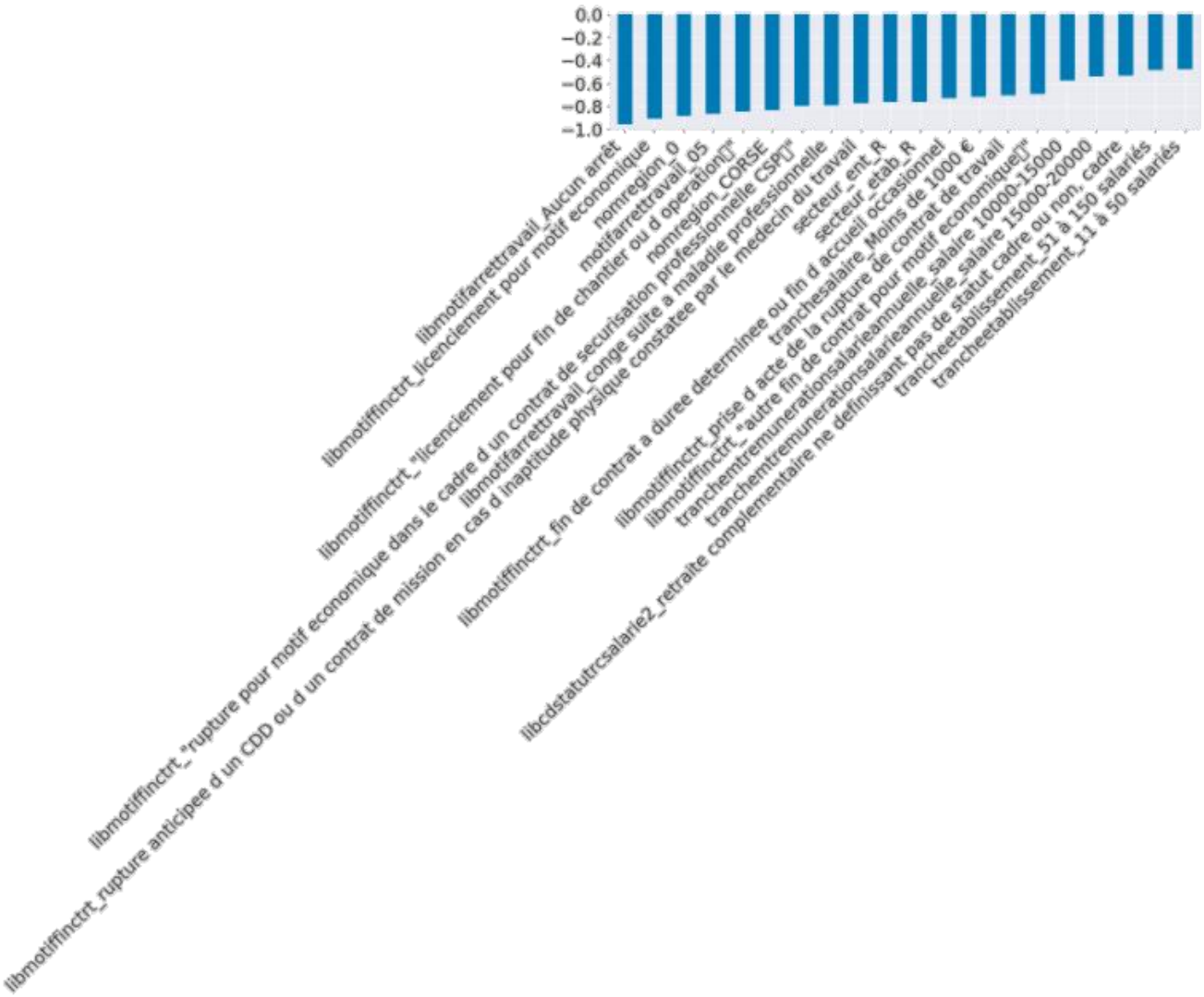
Coordonnées positives des variables par rapport au premier axe



Coordonnées positives des variables par rapport au second axe



Coordonnées négatives des variables par rapport au premier axe



Coordonnées négatives des variables par rapport au second axe