**Mémoire présenté le 8 juillet 2021 :**

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires**

Par : Minh Tu Pham

Titre : Construction de tables de turnover par application de l'approche d'apprentissage automatique dans l'évaluation des Indemnités de Fin de Carrière en norme IAS 19.

Confidentialité : ☒ NON ☐ (Durée : ☐ 1 an ☐ 2 ans)
*Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus*

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

Leila Aziri

*Membres présents du jury de l'ISFA*

Diana Dorobantu

*Entreprise :*
*Nom :* Aon France

*Signature :*

*Directeur de mémoire en entreprise :*
*Nom :* Yankel Benasuly

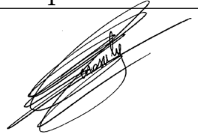
*Signature :*

*Invité :*
*Nom :* Stéphane Rebaudo

*Signature :*

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

Signature du responsable entreprise



Signature du candidat

# Remerciements

Je tiens tout d'abord à remercier Monsieur Yankel Benasuly - Responsable de l'offre Actuariat du pôle Retraite et Investissement d'Aon France, pour son encadrement, pour ses disponibilités et ses conseils précieux sur le sujet.

Je souhaite remercier également les autres managers de l'équipe Actuariat, Monsieur Stéphane Rebaudo et Madame Helena Niamké pour m'avoir encadré tout au long de cette étude, pour leurs disponibilités, leurs remarques et le suivi régulier.

Je remercie également Monsieur Cyril Manach et Monsieur Benjamin Lewis, les consultants de l'offre Actuariat pour leurs conseils précieux et leurs relectures avisées.

Je tiens à remercier particulièrement Monsieur Nabil Kazi-Tani, mon tuteur à l'Institut de Science Financière et d'Assurances (ISFA) d'avoir encadré ce mémoire. Merci également à l'ensemble des professeurs de l'ISFA pour les connaissances que j'ai apprises et sans lesquelles la rédaction de ce mémoire aurait été compromise.

Je tiens également à remercier ma famille pour m'avoir soutenu et aidé tout au long de mes études.

# Résumé

Mots clés : turnover, norme IAS 19, Indemnités de Fin de Carrière, machine learning, apprentissage automatique.

Ce mémoire traite de la problématique de l'hypothèse de taux de turnover (ou taux de rotation du personnel) dans le calcul des Indemnités de Fin de Carrière (IFC) d'une entreprise. Cette hypothèse est une des plus difficiles à établir car elle n'est pas clairement définie dans la norme IAS 19. Selon la norme IAS 19, il est donc nécessaire pour chaque entreprise de construire une table de turnover reflétant le plus fidèlement possible la réalité des taux de sorties observées au sein de cette entreprise et permettant de représenter les comportements sur la durée restante de l'engagement. L'hypothèse de turnover ayant un impact considérablement significatif sur l'engagement des avantages sociaux, les cabinets d'audit en France ont porté une attention particulière ces dernières années sur cette hypothèse, appuyée par une note rédigée en octobre 2018 de la Compagnie Nationale des Commissaires aux Comptes (CNCC) dans laquelle il est précisé que l'évaluation des engagements de retraite doit être effectuée en tenant compte des seules prévisions de démissions, à l'exclusion des licenciements et ruptures conventionnelles.

En pratique, les tables utilisées pour cette hypothèse sont construites en déterminant le taux de sortie à chaque âge pour la population concernée. Néanmoins, la modélisation du turnover est relativement difficile du fait de données parfois insuffisantes mais aussi du fait de la forte variabilité de cette hypothèse. De nombreux facteurs qui pourraient avoir des impacts significatifs sur le taux de démission d'un salarié n'ont pas été pris en compte comme par exemple : l'ancienneté, le salaire, la catégorie, le sexe, le statut matrimonial, etc. Ainsi, une approximation trop grossière des taux de sortie peut engendrer des erreurs significatives. Pour le lecteur et la bonne compréhension de ce mémoire, il est rappelé que plus le taux de rotation du personnel est élevé, plus l'engagement calculé est faible.

Ce mémoire présente une nouvelle approche de construction de tables de turnover ainsi que l'impact de celles-ci sur l'engagement. Le taux de turnover est déterminé au niveau individuel en fonction de ses propres caractéristiques. Chaque table de turnover est construite de la manière prospective pour chaque salarié, c'est-à-dire qu'elle tient compte des évolutions à venir des caractéristiques.

Cette nouvelle approche est réalisée à l'aide des techniques d'apprentissage automatique (ou de *machine learning*). Heureusement pour les entreprises, grâce à la rapidité des développements de l'intelligence artificielle, de la baisse des prix du stockage et de la puissance de calculs, les capacités de *machine learning* sont devenues de plus en plus accessibles (Shmueli, Patel, &

Bruce, 2010 ; Witten, Frank, Hall & Pal, 2016). Outre l'augmentation de la puissance de calculs et du stockage, le volume de données collectées et disponibles gratuitement par les entreprises a également considérablement augmenté (Goodfellow, Bengio & Courville, 2016 ; Shmueli et al., 2010). Les algorithmes d'apprentissage automatique se nourrissent de ces données. C'est ce qu'ils utilisent pour apprendre, comprendre des modèles et repérer des tendances.

L'objectif de cette étude est d'examiner les opportunités des techniques d'apprentissage automatique présentes aujourd'hui, et voir comment ces techniques peuvent aider à construire des tables de turnover qui reflètent les meilleures estimations de rotation du personnel au sein de l'entreprise. Cela conduit aux trois questions clés suivantes :

1. Quel algorithme d'apprentissage automatique est le plus approprié pour prévoir le turnover ?
2. Quels sont les prédicteurs importants pour déterminer le turnover ?
3. Comment construire des nouvelles tables de turnover ?

Pour répondre à ces questions, les analyses exploratoires et les traitements de données ont été réalisés avant de lancer les modèles d'apprentissage automatique. Les analyses de données nous permettent de mieux comprendre la base de données, d'anticiper des problèmes potentiels et de proposer les algorithmes d'apprentissage automatique appropriés. D'ailleurs, les techniques de traitement de données, telles que le regroupement des modalités et la numérisation pour les variables catégorielles, la standardisation et la transformation Box-Cox pour les variables numériques, ont pour but d'améliorer la performance des algorithmes.

L'ensemble des algorithmes d'apprentissage automatique qui ont été utilisés dans cette étude sont les suivants : la régression logistique et ses versions pénalisées, les modèles non-linéaires (K plus proches voisins, machines à vecteurs de support, réseau de neurones), l'arbre de décision et les modèles d'ensemble. Par ailleurs, les données s'exposent à un problème de déséquilibre car le taux de turnover moyen observé est seulement de 4,13 %. En effet, les techniques de ré-échantillonage ont été sollicitées afin de résoudre ce problème.

Parmi les modèles ci-dessus, le meilleur modèle, selon les critères d'évaluation de modèles, a été choisi pour construire des nouvelles tables de turnover. Dans cette étude, les deux types de tables de turnover seront proposés : les tables individuelles (i.e. à chaque salarié sa propre courbe de taux de turnover) et une table prospective pour l'ensemble des salariés de l'entreprise.

Le graphique suivant présente la table de turnover prospective donnée en fonction de l'âge et de l'année future.

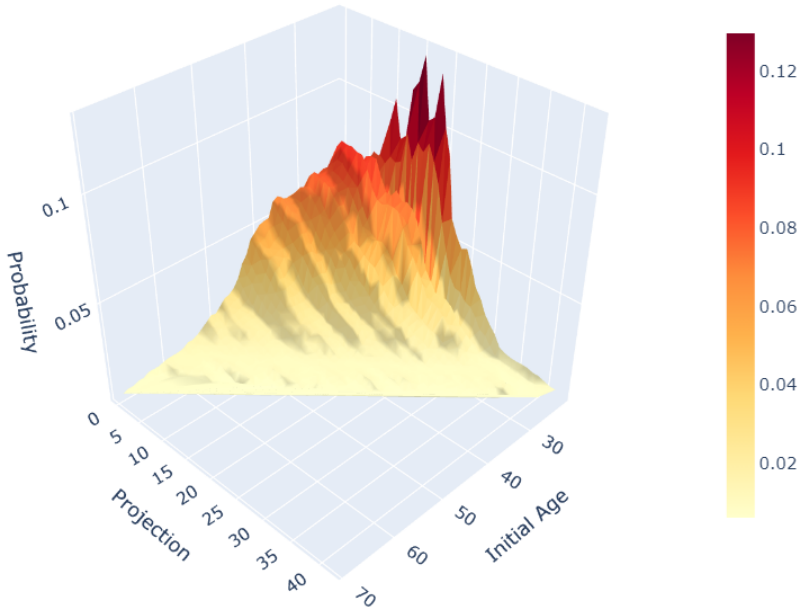



FIGURE 1 – Taux de turnover prospectifs

L'application des nouvelles tables de turnover entraîne des impacts importants sur l'engagement des IFC :

- Les tables individuelles : une baisse de 14 % sur l'engagement, soit environ 138 millions d'euros.
- La table prospective : une baisse de 17 % sur l'engagement, soit environ 160 millions d'euros.

Ce mémoire propose une nouvelle approche d'affinement du turnover et démontre que l'engagement est très sensible à cette hypothèse. L'actuaire doit veiller à calibrer les taux de rotation dans le calcul des passifs sociaux afin de se conformer aux exigences de la norme IAS 19.

# Abstract

Keywords : turnover, IAS 19, Retirement Indemnity, machine learning.

This thesis deals with the issue of the turnover assumption in the calculation of the retirement indemnity (IFC) for a company. This assumption is one of the most difficult assumptions to establish because it is not clearly defined in IAS 19. According to the IAS 19 standard, it is therefore appropriate for each company to design a turnover table reflecting as closely as possible the reality of the turnover rates observed within this company and allowing future behaviors to be represented over the remaining period of the liability. Since the assumption of turnover has a considerable impact on the liability of pension and social benefits plans, audit firms in France have paid particular attention over the recent years to this assumption. This is supported by a note issued in October 2018 from the Compagnie Nationale des Commissaires aux Comptes (CNCC) in which it is specified that the valuation of pension liabilities must be carried out taking into account the only forecasts of resignations, to the exclusion of redundancies and contractual terminations.

In practice, the tables used for this assumption are constructed by determining the turnover rate at each age for the population concerned. However, modeling turnover is relatively difficult due to sometimes insufficient data but also due to the high variability of this assumption. Several factors that could have significant impacts on an employee's resignation rate have not been taken into account, such as : seniority, salary, category, sex, marital status, etc. Thus, an over rough approximation of turnover rates can lead to significant errors. For the reader and the understanding of this thesis, it is recalled that the higher the employee turnover rate, the lower the calculated liability.

This thesis presents a new approach in constructing turnover tables as well as their impacts on the liability. The turnover rate is determined at the individual level based on their own characteristics. Each turnover table is constructed in a prospective manner for each employee, i.e. it takes into account future changes in characteristics.

This new approach is achieved by using machine learning techniques. Fortunately for companies, due to fast pace of developments in artificial intelligence, the decreasing prices of storage and computing power ; machine learning capabilities have become increasingly more accessible (Shmueli, Patel, & Bruce, 2010 ; Witten, Frank, Hall, & Pal, 2016). Besides the increase in computing power and storage, also the volume of data that companies collect and is freely available has drastically increased (Goodfellow, Bengio, & Courville, 2016 ; Shmueli et al., 2010). Machine learning algorithms feed on this data. It is what they use to learn, figure out patterns,

and spot trends.

The objective of this study is to examine apporpriate machine learning opportunities present today, and to see how machine learning techniques could help to construct turnover tables that reflect the best estimates of resignation rates within the company. This leads to the following three key questions :

1. What machine learning algorithm is most appropriate for predicting employee voluntary leave ?
2. What are the significant predictors for determining employee voluntary leave ?
3. How to contruct new turnover tables ?

To answer these questions, data exploratory analyzes and data processing were carried out before launching the machine learning models. Data analyzes allow us to better understand the database, anticipate potential problems, and propose appropriate machine learning algorithms. Moreover, data processing techniques, such as grouping of categories and numerization for categorical variables, standardization and Box-Cox transformation for numeric variables, aim to improve the performance of the algorithms.

Machine learning algorithms that were used in this study are as follows : logistic regression and its penalized versions, non-linear models (K nearest neighbors, support vector machines, neural network), decision tree and ensemble models. In addition, the data expose a problem of imbalance where the observed average turnover rate is only 4.13%. Therefore, resampling techniques have been tried in order to resolve this problem.

Among the above models, the best model, according to the model evaluation criteria, was chosen to construct new turnover tables. In this study, there are two types of turnover tables that will be proposed : individual turnover tables (i.e. each turnover rate curve specific to each employee) and a prospective turnover table for all of the company's employees.

The following graph will show the prospective turnover table which is based on age and future year.

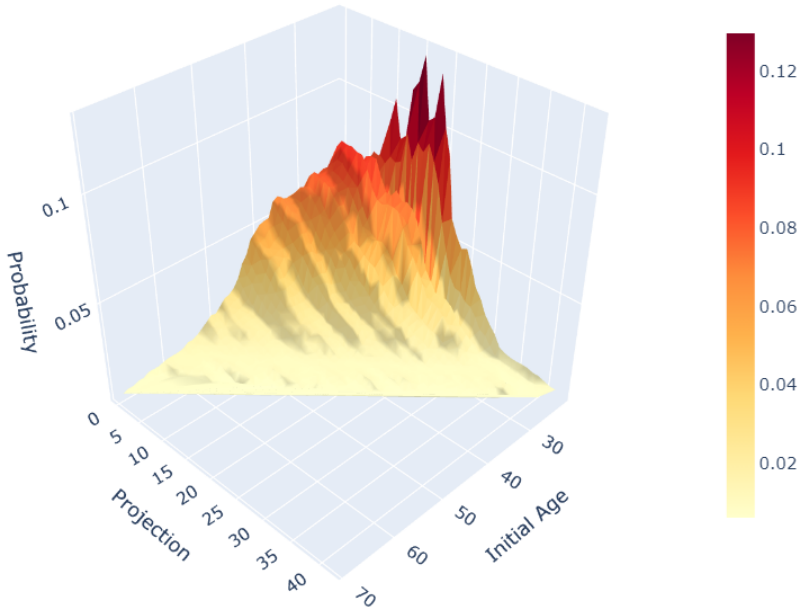



FIGURE 2 – Prospective turnover rates

The application of the new turnover tables has significant impacts on the liability of IFC :

- Individual turnover tables : a decrease of 14% on liability, or roughly 138 million euros.
- The prospective turnover table for all employees : a decrease of 17% on liability, or roughly 160 million euros.

This thesis proposes a new approach in refining turnover and shows that the liability is highly sensitive to this assumption. The actuary must pay attention to calibrate turnover rates in calculating social liabilities in order to comply with the IAS 19 requirements.

# Sommaire

# Table des figures

# Liste des tableaux

# Liste des sigles et acronymes

AGIRC = Association Générale des Institutions de Retraite de Cadres

ARRCO = Association pour le Régime de Retraite Complémentaire des Ouvriers

CNCC = Compagnie Nationale des Commissaires aux Comptes

DREES = Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques

IAS = International Accounting Standards

IASB = International Accounting Standards Board

IC = Interest Cost

IFC = Indemnités de Fins de Carrières

IFRS = International Financial Reporting Standards

NC = Normal Cost

OCI = Other Comprehensive Income

PASS = Plafond Annuel de la Sécurité Sociale

DBO = Defined Benefit Obligation

RH = Ressources Humaines

SC = Service Cost

SoRIE = Statement Of Recognized Income and Expense

VAP = Valeur Actuelle Probable

# Introduction

Aujourd'hui, l'évaluation des engagements au titre des passifs sociaux représente un réel enjeu pour les sociétés. Un passif social représente un engagement de l'entreprise envers ses salariés, il constitue pour cette dernière une dette dont le paiement est différé dans le temps. Il est donc indispensable pour l'entreprise de connaître la valeur réelle de ses engagements même si ceux-ci correspondent à des prestations futures non certaines.

Les engagements à évaluer peuvent être de différentes natures : les Indemnités de Fin de Carrière, les médailles du travail, les régimes supplémentaires de retraite, etc. Néanmoins, les IFC représentent l'essentiel des évaluations réalisées. En effet, cette prestation est généralement définie dans les conventions collectives nationales ou les accords d'entreprises et constituent des obligations conventionnelles pour la quasi-totalité des entreprises.

Pour évaluer ces engagements, qui sont « la valorisation d'un salaire ou gratification différés, actualisés et probabilisés », un actuaire procède en deux temps. Dans un premier temps, il doit obtenir de l'entreprise une base de données contenant les paramètres nécessaires pour les évaluations tels que le genre, l'âge, l'ancienneté, la catégorie, et le salaire de chaque salarié éligible. Il va réaliser les analyses sur ces données afin de s'assurer de leur cohérence. Dans un second temps, il formule des hypothèses actuarielles nécessaires au calcul des engagements, tout en étant en conformité avec les normes locales ou internationales. Les hypothèses retenues sont à la fois la décision de l'entreprise et de son actuaire. Parmi les hypothèses majeures à établir, on distingue deux natures : les hypothèses financières qui sont : le taux d'actualisation, l'inflation long terme et le taux de revalorisation des salaires, et les hypothèses démographiques qui permettent de considérer les spécificités propres des bénéficiaires du régime à l'égard de certains critères tels que l'âge de départ à la retraite, la mortalité, ou le taux de turnover. Le choix de ces hypothèses est fondamental puisque chaque paramètre a un impact direct sur l'engagement à évaluer.

L'évaluation et la comptabilisation des avantages au personnel sont définies, depuis le 1er janvier 2005, dans la norme IAS 19. Cette norme impose aux entreprises de constituer des provisions pour tous les avantages octroyés au personnel. Ces avantages constituent généralement d'importantes sommes et ce qui nécessite pour les entreprises d'anticiper leur paiement et donc de les provisionner. La norme IAS 19 permet ainsi une standardisation des méthodes de calcul et de la comptabilisation pour les engagements sociaux et indique des règles précises quant à leur traitement. De plus, la norme IAS 19 préconise un certain nombre de règles pour le choix des hypothèses actuarielles. Les choix opérés doivent être justifiés et conformes aux indications prévues par la norme. Le turnover constitue une des hypothèses actuarielles des plus délicates à

établir. En effet, contrairement à d'autres hypothèses comme le taux d'actualisation, la norme IAS 19 ne donne pas d'indications précises quant à la détermination des tables de turnover. Il convient donc de déterminer pour chaque entreprise une table de turnover qui reflète le plus finement possible les taux de sortie chaque année. Généralement, ces taux de sortie sont déterminés à partir des données transmises par l'entreprise, en se basant sur les deux ou trois années précédentes. Cette méthode permet de construire une table de turnover en fonction de l'âge et/ou de la catégorie. Néanmoins, de nombreux écarts sont observés entre le turnover choisi dans les hypothèses et le turnover réel, donnant lieu à la constatation systématique d'écarts actuariels. Cela s'expliquerait par le fait que de nombreux facteurs ayant un impact important sur le turnover n'ont pas été considérés dans la construction de la table de turnover, tels que le salaire, le statut matrimonial, le sexe, la zone géographique, etc.

L'objectif de ce mémoire est donc de donner une nouvelle approche en déterminant le taux de turnover au niveau individuel avec l'aide des modèles d'apprentissage automatique et d'examiner les impacts de nouvelles tables de turnover sur l'engagement. Les taux de turnover estimés sont propres à chaque salarié tenant en compte l'ensemble des caractéristiques de ce salarié.

Ce mémoire, dont le sujet porte sur « Construction de tables de turnover par application de l'approche d'apprentissage automatique dans l'évaluation des Indemnités de Fin de Carrière en norme IAS 19 », est structuré de cinq chapitres comme les suivants :

- Chapitre 1 : il expose tout d'abord la présentation sur le cadre de l'étude : la système de retraite et des avantages sociaux en France et la norme IAS 19. Les indemnités de fin de carrière et leurs calculs actuariels sont également présentés dans ce premier chapitre.
- Chapitre 2 : il présente le contexte et apporte plus de détails sur les motivations de la recherche d'une nouvelle approche de construction de la table de turnover. Il introduit ainsi le lecteur à un modèle de référence des évaluations des IFC dont les résultats servent comme les valeurs de référence à comparer avec ceux obtenus par des nouvelles tables de turnover.
- Chapitre 3 : il aborde une étape essentielle de l'étude sur l'ingénierie des données. Les données seront explorées and analysées en détail. En se basant sur les résultats des analyses, l'ensemble des techniques de traitement des données seront proposées.
- Chapitre 4 : il résume les idées principales des algorithmes d'apprentissage automatique utilisés dans cette étude ainsi que les métriques d'évaluation de modèles.
- Chapitre 5 : ce dernier chapitre fournit les résultats de l'étude et la construction de nouvelles tables de turnover. Ensuite, l'impact du passage à la nouvelle approche de construction de la courbe des taux de turnover sera exposé. Pour finir, la dernière section du chapitre résumera les limites auxquelles nous nous confrontons dans l'étude.

# Chapitre 1

# Cadre de l'étude

Ce premier chapitre constitue une introduction du cadre de l'étude en présentant premièrement le système de retraite et avantages sociaux en France, deuxièmement la norme IAS 19 et enfin les indemnités de fin de carrière.

## 1.1 La retraite et les avantages sociaux en France

Le système de retraite et avantages sociaux français est le fruit d'une histoire longue et complexe, ce qui explique la diversité des régimes et des caisses. Le régime des Indemnités de Fin de Carrière fait partie de ce système. Ainsi, il est intéressant de comprendre, en premier lieu, l'environnement qui encadre l'étude.

### 1.1.1 La retraite en France

Le système de retraite français est basé essentiellement sur le principe de la répartition, fondé sur la solidarité inter-générationnelle, et accessoirement par capitalisation.

**Principes généraux :**

On distingue deux méthodes de financement :

- **Le financement par répartition** : Le principe est simple et bien connu. Une caisse reçoit les cotisations versées par les personnes actives et les utilise pour payer les pensions des retraités. Les comptes sont équilibrés par année civile. Ce que la caisse touche une année donnée est distribué la même année. Ce principe de répartition repose donc sur une forte solidarité entre les actifs et les retraités, entre les plus jeunes et les plus âgés. On parle de solidarité intergénérationnelle.
- **Le financement par capitalisation** : La logique est différente. Les actifs d'aujourd'hui épargnent en vue de leur propre retraite dans un cadre individuel ou collectif. Les cotisations font l'objet de placements financiers ou immobiliers, dont le rendement, qui conditionne le niveau des pensions futures, dépend essentiellement des performances du marché sur le long terme, et donc des taux d'intérêts.

A ces deux modes de financement s'appliquent deux modes de gestion : à prestations définies (Defined Benefits) ou à cotisations (Defined Contributions).

- **Régimes à prestations définies** : L'entreprise s'engage sur le niveau de prestations (par exemple : régime de retraite avec une pension additive égale à 10 % du dernier salaire). Elle doit organiser le financement (en interne ou en externe) pour être en mesure d'honorer le paiement des prestations.
- **Régimes à cotisations définies** : L'entreprise s'engage sur un niveau de financement et non de prestations (par exemple : régime de retraite avec une cotisation de 4% des salaires). Les contrats de ce type de régime n'engendrent pas d'engagement pour l'employeur, le risque financier est supporté par le salarié. La comptabilisation des cotisations est en charge de l'exercice.

Ainsi, seuls les régimes à prestations définies doivent faire l'objet d'une évaluation actuarielle.

### Chiffres et dates clés

Le graphique suivant résume quelques statistiques de la retraite en France dans le rapport publié par la DREES (l'édition 2020) :

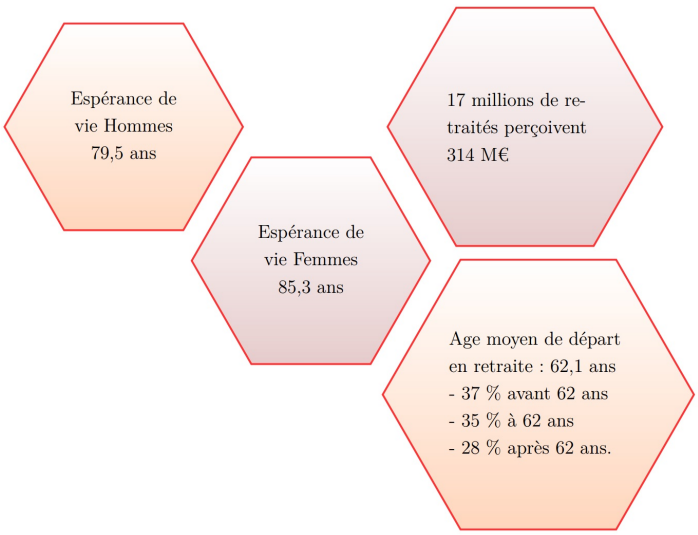



FIGURE 1.1 – Chiffres de la retraite en France en fin 2020

Panorama des réformes et des grandes nouveautés du système de retraite en France depuis sa création :

**1945** : Création de la Sécurité sociale.

**1947** : Création de l'AGIRC dédiée aux cadres de l'industrie et du commerce.

**1961** : Création de l'ARRCO dédiée aux non cadres.

**1972** : Généralisation de la retraite complémentaire à tous les salariés

**1974** : Affiliation obligatoire des cadres à l'ARRCO (en plus de celle à l'AGIRC).

**1982** : Ouverture possible des droits de retraite dès l'âge de 60 ans et accord pour la liquidation des retraites complémentaires avant 65 ans sans abattement (si les conditions du taux plein dans le régime général sont remplies).

**1993** : Réforme Balladur. Le nombre de trimestres requis pour la retraite à taux plein passe de 150 à 160 ans. Calcul du salaire annuel moyen sur les 25 meilleures années de salaire et mise en place de l'indexation des pensions sur l'indice INSEE des prix à la consommation.

**2003** : Réforme Fillon. Augmentation du nombre de trimestres requis pour la retraite à taux plein (de 160 à 164). Création des mesures de surcote et de cumul emploi-retraite. Taxation des dispositifs pré-retraite et promotion à l'embauche des séniors.

**2010** : Loi 9 Novembre 2010. Recul progressif de l'âge d'ouverture des droits (de 60 à 62 ans) et de l'âge de taux plein (de 65 à 67 ans). Allongement des durées de cotisation pour l'obtention du taux plein à 165 et 166 trimestres. Encouragement des dispositifs d'épargne retraite.

**2013** : Loi de Réforme des retraites. Allongement du nombre de trimestres requis pour la retraite à taux plein à 172 trimestres. Mise en place d'un compte pénibilité à compter de 2015.

**2019** : Fusion AGIRC-ARRCO. Les 2 caisses de retraite complémentaire des salariés du privé – l'AGIRC et l'ARRCO – ont prévu de fusionner, dans un accord signé le 30 octobre 2015 par les partenaires sociaux, pour éviter la faillite du système

**Organisation du système**

Le système de retraite français s'est mis en place progressivement depuis 1945, pour l'ensemble des salariés du privé. Aujourd'hui, la retraite est articulée en France autour de 3 piliers qui sont les régimes obligatoires, les régimes supplémentaires d'entreprises et les régimes de retraites individuels.

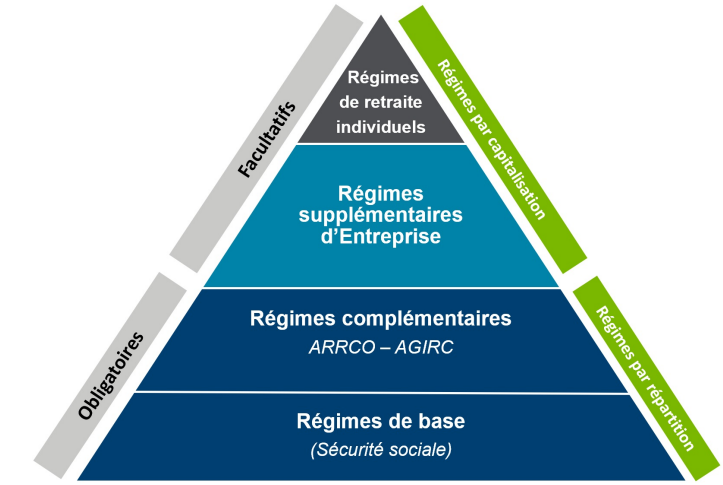



FIGURE 1.2 – Le système de retraite en France

Les régimes obligatoires comptent le régime de base et le régime complémentaire. Comme évoqué brièvement plus haut, ce premier pilier constitue le coeur du système de retraite français. Du fait que celui-ci fasse appel à la solidarité intergénérationnelle et au financement par répartition, il est donc assuré par le salarié et le cas échéant, par son employeur.

Les régimes de retraite supplémentaires viennent s'ajouter aux régimes de base et aux régimes complémentaires. Ils constituent le deuxième pilier du système de retraite. On peut citer le régime supplémentaire de retraite à prestations définies où le montant des futures rentes est défini à l'avance. Le régime supplémentaire de retraite à cotisations définies prévoit un taux de cotisation négocié entre l'entreprise et les représentants syndicaux.

Le dernier pilier des produits d'épargne retraite individuel est géré en capitalisation, car l'adhésion à ce régime résulte d'une décision individuelle. Parmi ces produits, on compte le PERP (Plan d'épargne retraite populaire, accessible à tous), les contrats d'épargne-retraite Madelin (pour les indépendants), la Préfon pour les fonctionnaires.

Cette organisation détient une forte dimension sociale qui se dénote notamment par la reconnaissance d'un minimum vieillesse ainsi que d'un minimum social.

Les couvertures sociales diffèrent en fonction des types de profession tels que les salariés (du privé, agricole, des régimes spéciaux, de l'Etat, etc.) et les non-salariés (exploitants agricoles, commerçants et industriels, artisans et professions libérales). Cette diversité de sous-régimes a pour conséquence un nombre élevé de régimes et de caisses illustrés dans le schéma en **Annexe A**.

### 1.1.2 Les avantages sociaux en France

Les avantages sociaux constituent une part de rémunération s'ajoutant au salaire direct, sous toute forme de paiements, d'indemnités et d'autres services accordés aux salariés, par les entreprises ou par l'Etat.

Les avantages sociaux comprennent les régimes de complémentaire santé, de prévoyance, des congés payés, des congés de maladie, des tickets restaurant, le remboursement de certains frais de scolarité, le remboursement (en tout ou en partie) des frais reliés à l'utilisation des transports en commun, etc.

En particulier, la plupart des entreprises françaises offrent aux travailleurs le type standard d'avantages sociaux suivant :

- Régime d'épargne salariale - retraite : Plusieurs entreprises offrent à leurs employés un régime enregistré d'épargne-retraite (REER) collectif à adhésion facultative ou un régime semblable. Dans certains cas, l'employeur y contribue également selon un pourcentage du montant versé par l'employé.

Le panorama de l'épargne retraite est résumé dans le graphique suivant :

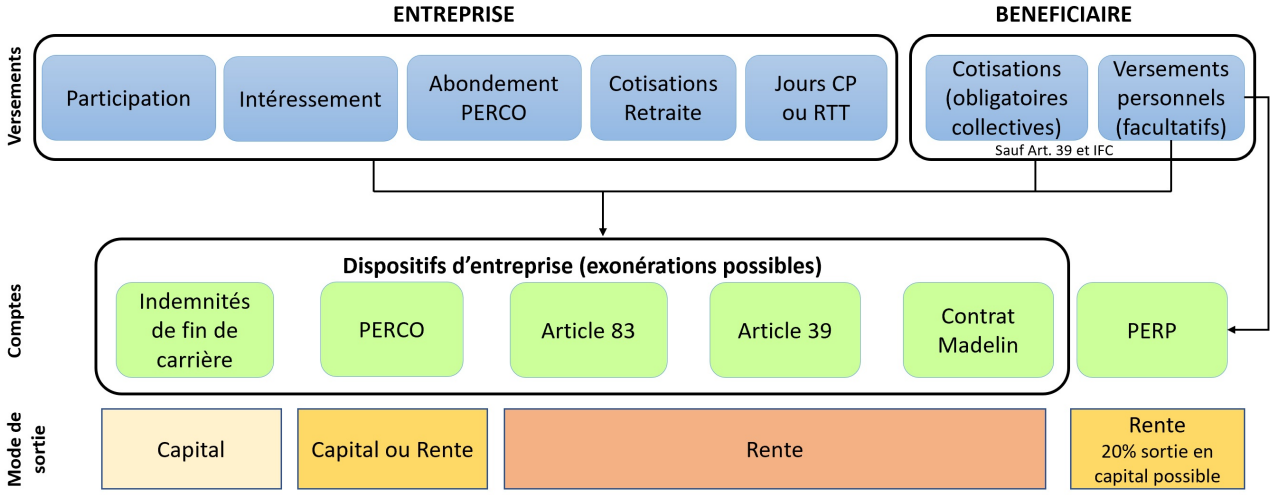



FIGURE 1.3 – Panorama de l'épargne retraite entreprise et individuelle

Nous distinguons sur ce schéma :

- La première ligne qui représente l'ensemble des versements possibles, soit par l'entreprise, soit par le bénéficiaire.
- La deuxième ligne qui résume les différents comptes sur lesquels les versements peuvent

être réalisés. Ces comptes sont donc « alimentés » par l'entreprise et/ou le bénéficiaire. Ils appartiennent exclusivement au bénéficiaire.

- La troisième ligne qui précise pour chaque compte la sortie possible, en rente ou en capital.

**Les différents contrats d'avantages sociaux en entreprise**

**Les contrats à prestations définies :**

- Les contrats d'indemnités de fin de carrière (IFC) : Lorsqu'un salarié quitte l'entreprise au moment de son départ en retraite, l'employeur est tenu de lui verser une IFC. Le montant de cette indemnité est défini par la loi, la convention collective applicable à l'entreprise, l'accord d'entreprise ou le contrat de travail du salarié concerné.
- Les contrats « Articles 39 », souvent appelés « retraites chapeaux » : ces régimes mis en place dans certaines entreprises concernent généralement une catégorie spécifique de salariés (souvent les cadres et cadres supérieurs). Ce sont des régimes à « prestations définies » : l'entreprise s'engage dès le départ sur le montant de revenu supplémentaire qui sera versé par la suite à chaque bénéficiaire. En revanche, le salarié doit, le plus souvent, rester dans l'entreprise jusqu'à la retraite afin de pouvoir en bénéficier.

**Les contrats à cotisations définies :**

- Les contrats relevant de l'article 82 du Code Général des impôts (CGI) : Abondés exclusivement par l'employeur. Ils permettent aux salariés d'obtenir le versement d'une rente ou d'un capital. Les cotisations sont imposables au titre de l'impôt sur le revenu car elles sont considérées comme un « sur-salaire ».
- Les contrats « PER Entreprises (ou Article 83) » : Les cotisations obligatoires sont fixées à l'avance, à un niveau constant, et permettent d'obtenir une rente à la retraite. Ces versements peuvent être complétés par des versements volontaires à l'initiative des salariés adhérents. Le « PER Entreprises » est avantageux fiscalement pendant la phase de constitution de l'épargne-retraite, puisque les versements volontaires sont déductibles du revenu imposable (dans certaines limites). Une fois le salarié à la retraite, la rente « PER Entreprises » s'ajoute aux pensions des régimes obligatoires, et bénéficie du même régime fiscal et social.
- Le PERCO (Plan d'Epargne pour la Retraite Collectif) : à l'instar des autres types de contrats, il s'agit également d'un produit d'épargne retraite mis en place dans le cadre de l'entreprise. Ce contrat est alimenté par les versements du salarié et les éventuels abondements de l'employeur. Ces sommes resteront acquises en cas de départ de l'entreprise. De plus, si la nouvelle entreprise dispose également d'un PERCO, il est possible de transférer directement son épargne sur ce nouveau plan.

## 1.2 La norme IAS 19

Cette section présente l'historique la norme internationale IAS 19 qui décrit les exigences comptables relatives aux avantages du personnel, puis explicite les méthodes d'évaluation et de comptabilisation dans le cadre de l'IAS 19. Une brève comparaison entre la norme IAS 19 et la norme française sera présentée à la fin du chapitre.

### 1.2.1 Présentation de la norme

L'IASB (International Accounting Standards Board) a été créé en 1973 par les instituts comptables de 9 pays, dont la France et s'intitulait jusqu'en 2001 l'International Accounting Standards Committee. Il a pour objectif d'élaborer et de publier des normes internationales d'information financière pour la présentation des états financiers, ainsi que de promouvoir leur utilisation et leur généralisation à l'échelle mondiale. Ces normes étaient appelées jusqu'en 2011 les normes IAS (International Accounting Standards) et sont dorénavant appelées IFRS (International Financial Reporting Standards). Elles ont pour principal but d'améliorer la sécurité financière.

Ces nouvelles normes d'information financière marquent une évolution de la comptabilité vers une approche économique et un souci d'évaluer au mieux la performance financière des entreprises. La primauté de la réalité économique sur l'apparence juridique, l'évaluation à la juste valeur (« fair value »), le recours à l'actualisation ainsi que l'exigence d'une information très complète dans l'annexe constituent les principales caractéristiques de ces nouvelles normes.

La norme IAS 19 « Avantages du personnel » est entièrement consacrée au traitement des engagements sociaux. Elle impose notamment les méthodes d'évaluation et de comptabilisation à appliquer mais également la liste des informations à publier en annexes des comptes. Cette norme est complétée par la Norme IFRS 2 « Paiements en actions » qui traite notamment de la comptabilisation des plans de stock-option.

La première version de la norme IAS 19 a été publiée par l'IASB en février 1998 et homologuée par le règlement CE n° 1725/2003 du 29 septembre 2003. Puis, le 3 novembre 2008, la CE a regroupé les normes et interprétations, en un seul texte (le règlement CE n° 1126/2008), adoptées intégralement par la CE le 15 octobre 2008.

Suite à des amendements apportés à IAS 19 en juin 2011 par l'IASB concernant les régimes à prestations définies, une nouvelle version « révisée » d'IAS 19 a été publiée par l'IASB qui

s'applique obligatoirement aux périodes ouvertes à compter du 1er janvier 2013. Cette nouvelle version d'IAS 19 a été adoptée le 5 juin 2012 par l'UE (règlement UE n° 475/2012), avec une date d'application obligatoire pour les périodes ouvertes à compter du 1er janvier 2013, une application anticipée étant autorisée à l'instar de l'IASB.

### 1.2.2 Champs d'application

La norme IAS 19 doit être appliquée pour la comptabilisation de tous les avantages du personnel, sauf ceux auxquels s'applique la norme IFRS 2 « Paiement en actions ».

Les avantages du personnel sont classés en 4 catégories distinctes, pour chacune d'entre elles la norme IAS 19 révisée prescrit des dispositions spécifiques :

- les avantages à court terme,
- les avantages postérieurs à l'emploi,
- les autres avantages à long terme,
- les indemnités de cessation d'emploi.

1. Les **avantages à court terme** sont les avantages du personnel (autres que les indemnités de cessation d'emploi) dont le règlement intégral est attendu dans les douze mois qui suivent la clôture de l'exercice au cours duquel les membres du personnel ont rendu les services correspondants.

A titre d'exemples : les salaires et cotisations de sécurité sociale, les congés annuels payés et les congés de maladie payés, l'intéressement et les primes, les avantages en nature comme l'assistance médicale, le logement, la voiture et les autres biens ou services gratuits ou subventionnés dont bénéficient les membres du personnel en activité.

2. Les **avantages postérieurs à l'emploi** sont les avantages du personnel (autres que les indemnités de cessation d'emploi et les avantages à court terme) payables après la cessation de l'emploi du membre du personnel. On distingue deux types de régimes d'avantages postérieurs à l'emploi :

- les régimes à cotisations définies ;
- les régimes à prestations définies.

La plupart sont des prestations de retraite. Pour les régimes à cotisations définies, l'entité doit comptabiliser les coûts annuels. En revanche, pour les régimes à prestations définies, l'entité doit réaliser une évaluation actuarielle, fondée sur des hypothèses et des méthodes de projection, afin de mesurer son obligation et calculer ses charges annuelles.

A titre d'exemples : ***les indemnités de fin de carrière***, les autres prestations de prévoyance telles que des couvertures médicales, des assurances-vie ou des assurances-décès postérieures à l'emploi ou encore des avantages en nature maintenus pour les retraités.

3. Les **autres avantages à long terme** sont tous les avantages du personnel autres que les avantages à court terme, les avantages postérieurs à l'emploi et les indemnités de cessation d'emploi. Les autres avantages à long terme sont les avantages dont le règlement intégral est attendu au-delà de douze mois suivant la clôture de l'exercice au cours duquel les membres du personnel ont rendu les services correspondants.

A titre d'exemples : les absences de longue durée rémunérées, comme les congés liés à l'ancienneté ou les congés sabbatiques, les primes d'ancienneté et autres avantages liés à l'ancienneté, les indemnités pour invalidité de longue durée, l'intéressement, les primes et la rémunération différée dont le règlement intégral est attendu plus de douze mois suivant la clôture de l'exercice au cours duquel les membres du personnel ont rendu les services correspondants.

4. Les **indemnités de cessation d'emploi** sont les avantages du personnel fournis en contrepartie de la cessation d'emploi d'un membre du personnel résultant : (i) soit de la décision de l'entité de mettre fin à l'emploi du membre du personnel avant l'âge normal de départ en retraite, (ii) soit de la décision du membre du personnel d'accepter une offre d'indemnités en échange de la cessation de son emploi.

A titre d'exemples : les indemnités de licenciement, les indemnités versées dans le cadre de plans de départ en préretraite ou de plans de départ volontaire - lorsqu'elles ne sont pas qualifiées d'avantages postérieurs à l'emploi, les indemnités de départ transactionnelles.

### 1.2.3 Méthodes d'évaluation en norme IAS 19

Cette partie concerne plus particulièrement les avantages postérieurs à l'emploi et les autres avantages à long terme, sous la norme IAS 19 révisée (2013). On parle de l'engagement ou « Defined Benefit Obligation » (DBO) qui représente le passif social à comptabiliser. La norme IAS 19 impose la méthode à retenir pour estimer la DBO. Il s'agit de la méthode des « Unités de Crédit Projetées » (Projected Unit Credit Method). On distingue alors deux variantes à celle-ci :

1. Unités de Crédit Projetées « escalier » :

Cette méthode est appelée méthode « escalier » parce qu'elle se base sur les droits acquis

à la date de calcul. Ainsi, la DBO est calculée comme étant égale à la VAP (Valeur Actuelle Probable) des prestations prévues par le régime multipliée par le prorata entre les droits acquis à la date d'évaluation et les droits au terme. La formule de calcul de la DBO est la suivante :

$$Engagement\ (DBO) = VAP \frac{Droits\ acquis}{Droits\ au\ terme} \tag{1.1}$$

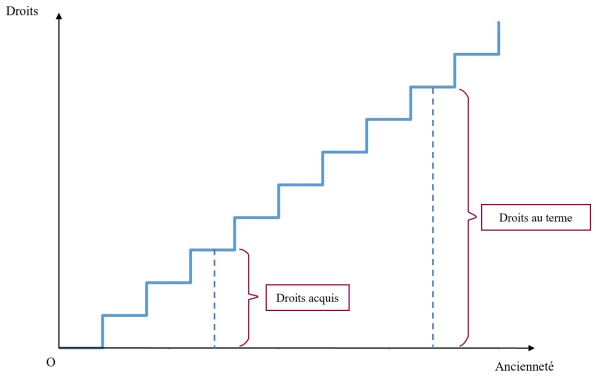



FIGURE 1.4 – Unités de Crédit Projetées « escalier »

C'est la méthode de référence préconisée par la norme IAS 19 pour la détermination de la DBO. Néanmoins, cette méthode présente l'inconvénient d'être mal adaptée au contexte français (notamment pour les Médailles du Travail) et ce n'est pas la méthode que l'on utilisera dans la suite de l'étude.

2. Unités de Crédit Projetées avec « Services proratés » :

Cette méthode calcule les engagements au prorata de l'ancienneté actuelle sur l'ancienneté au terme :

$$Engagement\ (DBO) = VAP \frac{Ancienneté\ actuelle}{Ancienneté\ au\ terme} \tag{1.2}$$

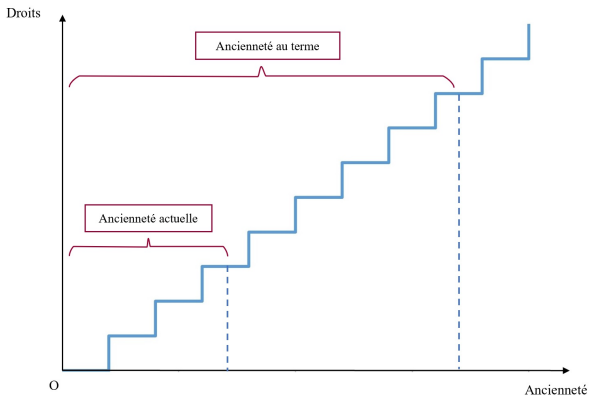



FIGURE 1.5 – Unités de Crédit Projetées avec « Services proratés »

Cette méthode prospective est plus adaptée au contexte français (notamment pour les IFC, les régimes de retraite « chapeau » et surtout pour les Médailles Du Travail...). Elle est donc utilisée en pratique chez Aon pour calculer les engagements sociaux.

Les calculs de la VAP vont être présentés plus précisément dans la section suivante.

### 1.2.4 Méthodes de comptabilisation en norme IAS 19

La comptabilisation des régimes à prestations définies est complexe en raison de la difficulté d'évaluation du passif. La norme IAS 19 encourage donc les entreprises à faire appel à un actuaire pour évaluer les engagements sociaux. Dans cette partie et dans les études de ce mémoire, conformément à la norme IAS révisée, nous allons présenter les méthodes de comptabilisation pour les avantages postérieurs à l'emploi et utiliser **la méthode SoRIE *(Statement of Recognition of Income and Expense)*** : Tous les écarts actuariels sont reconnus immédiatement dans l'année en autres éléments du résultat global.

Le montant comptabilisé au passif au titre des prestations définies se détermine comme suit :

Valeur actuelle de l'obligation à la fin de la période de reporting

– Juste valeur à la date de clôture des actifs du régime

= **Passif (ou provision) au titre des prestations définies**

Où la valeur actuelle de l'obligation (DBO) à la fin de l'exercice est calculée par :

**Engagement DBO au 31/12/N-1**

+ Coût des services de l'exercice N

+ Coût de l'actualisation de l'exercice N

– Prestations payées durant l'exercice N

+/– Variations de périmètre (acquisition, transferts,...)

+/– Pertes et (gains) actuariels

= **Engagement DBO au 31/12/N**

La graphique suivant montre le calcul de la provision sous la méthode SoRIE :

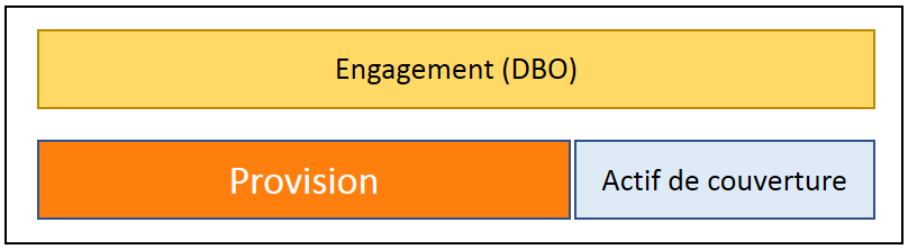



FIGURE 1.6 – La provision selon la méthode SoRIE

L'entreprise doit comptabiliser en **compte de résultat**, la somme des éléments suivants :

- le coût des services rendus au cours de l'exercice ;
- le coût des services passés (profits ou pertes relatifs à une modification ou une réduction de régime) ;
- le profit (ou la perte) résultant d'une liquidation de régime ;
- les intérêts nets sur le passif (l'actif) net au titre des prestations définies.

L'entreprise doit comptabiliser en **autres éléments du résultat global** (*Other Comprehensive Income (OCI)*) les réévaluations du passif (de l'actif) net au titre des prestations définies, elles correspondent à la somme des éléments suivants :

- les écarts actuariels liés à l'évaluation du passif et de l'actif ;
- la variation, le cas échéant, de l'effet du plafonnement de l'actif (à l'exclusion des montants pris en compte dans le calcul des intérêts nets sur le passif (l'actif) net).

Les éléments ci-dessus vont être précisés comme suivants :

**Current service cost** : Le « current service cost » est défini par la norme comme étant « le coût des services rendus au cours de l'exercice », et les éléments qui le composent sont identiques à ceux du DBO. Cette composante de la charge représente la charge d'une année supplémentaire de service.

**Coûts de services passés** : Le coût des services passés désigne la variation de la valeur actualisée de l'obligation pour les services rendus par les membres du personnel au cours de périodes antérieures, qui résulte :

- d'un Plan amendement : Effet de l'adoption ou de la modification des droits d'un régime sur l'obligation
  - ⊕ Changement du barème des droits (mise à jour de la convention collective par exemple).

⊕ Changement de la formule de calcul de la rente (plafonnement,limitation,...).

- d'une Réduction de régime : Curtailment
  - ⊕ Effet de la réduction significative du nombre de bénéficiaire du régime (licenciement économique par exemple).
  - ⊕ Lorsque l'entreprise change les termes du régime de sorte qu'un nombre significatif de bénéficiaires n'y a plus droit.

Selon l'IAS 19, les coûts des services passées doivent être comptabilisés immédiatement et en intégralité en compte de résultat l'année de survenance de l'événement.

**Liquidation d'un régime** : Il y a liquidation d'un régime lorsqu'une entreprise conclut une transaction éliminant toute obligation juridique ou implicite pour tout ou partie des prestations prévues. A titre d'exemple, il y a liquidation si l'entreprise règle aux bénéficiaires du régime une somme forfaitaire en échange de leurs droits à prestations. Ainsi, l'entreprise doit comptabiliser les pertes et profits au titre de la réduction ou de la liquidation d'un régime au moment où cette dernière se produit.

**Intérêt net** : L'intérêt net afférent à un régime à prestations définies se définit comme la charge d'intérêt (ou le produit) déterminé par application du taux d'actualisation de la dette à l'engagement net (déficit ou excédent du plan).

**Ecarts actuariels** : Les écarts actuariels sont les variations de la valeur actualisée de l'obligation au titre des prestations définies qui résultent :

- des ajustements liés à l'expérience : l'effet des écarts entre les hypothèses actuarielles antérieures et ce qui s'est effectivement produit ;
- des ajustements liés au changement d'hypothèses actuarielles : par exemple du taux d'actualisation, de turnover ou de la table de mortalité.

### 1.2.5 Comparaison entre la norme IAS 19 et la norme française (French GAAP)

Les différences entre les comptabilisations French Gaap et l'IAS 19 sont à appréhender étant donné que certaines entreprises fonctionnent encore suivant les normes françaises exclusivement. Le tableau ci-dessous regroupe les principales divergences pour les régimes postérieurs à l'emploi :

| | | IAS 19 Révisé | | French GAAP | |
|---|---|---|---|---|---|
| | | P&L | OCI | P&L | Corridor |
| | Defined Benefit Obligation BOY | | | | |
| | Current Service Cost | x | | x | |
| | Interest Cost | x | | x | |
| | Employee Contributions | | | | |
| **(1)** | **Plan amendment** | x | | x | Off-balance-sheet |
| **(1)** | **Curtailment** | x | | x | Off-balance-sheet |
| **(1)** | **Settlement** | x | | x | Off-balance-sheet |
| | Benefits paid | | | | |
| **(2)** | **Actuarial Gains / Losses** | | x | x | Off-balance-sheet |
| | Defined Benefit Obligation EOY | | | | |
| | | | | | |
| | Fair Value Asset BOY | | | | |
| | Employer cash out | | | | |
| | Expected return | x | | x | |
| | Employee Contributions | | | | |
| | Settlement | x | | x | |
| | Benefits paid | | | | |
| **(2)** | **Actuarial Gains / Losses** | | x | x | Off-balance-sheet |
| | Fair Value Asset EOY | | | | |

(1) Reconnaissance immédiate en P&L (Profit & Loss) en IAS 19 versus étalement possible en P&L et stock PSC (Prior Service Cost) en hors-bilan en French GAAP.

(2) Reconnaissance immédiate en OCI en IAS 19 versus amortissement possible en P&L et stock G/L (Gain/Loss) en hors-bilan avec corridor en French GAAP.

Méthode du corridor : Dans le cas ou le stock d'écart actuariel dépasse 10% de l'engagement, amortissement dans la charge de l'année N+1 de la partie excédentaire sur la durée de vie résiduelle du plan.

## 1.3 Les Indemnités de Fin de Carrière

Cette section vise à présenter les Indemnités de Fin de Carrière (IFC) et son calcul actuariel qui servent pour les calculs du modèle de référence dans le chapitre suivant.

L'indemnité de fin de carrière (IFC) est un régime à prestations définies qui garantit au salarié le versement d'une indemnité sous forme de capital lors de son départ à la retraite.

Conformément à la loi de mensualisation du 19 janvier 1978, cette indemnité est définie dans la convention collective nationale ou l'accord d'entreprise, ce dernier devant alors être plus avantageux. La prestation versée au titre de l'IFC, déterminée par l'ancienneté du salarié

au sein de l'entreprise, est le plus souvent basée sur le salaire de fin de carrière, mais peut parfois être définie par rapport à un forfait ou au PASS (Plafond Annuel de la Sécurité Sociale).

### 1.3.1 Méthode de calcul de l'IFC

La méthode de calcul de l'engagement des IFC dans le cadre de la norme IAS 19 est la méthode de « Unités de Crédit Projetées avec Services proratés » qui a été présentée dans la section 1.2.3 :

$$Engagement\ (DBO) = VAP \frac{Ancienneté\ actuelle}{Ancienneté\ au\ terme} \tag{1.3}$$

### 1.3.2 Calcul de la valeur actuelle probable (VAP)

La VAP représente la totalité des engagements à la date de l'évaluation, avec projection des salaires au terme.

La méthode de détermination de la VAP exige que les calculs soient effectués tête par tête. L'engagement global de l'entreprise envers ses salariés est alors donné par la somme des résultats individuels.

La valeur actuelle probable de l'engagement de l'entreprise envers un salarié à l'âge x à la fin de l'année n est égale à :

$$VAP = \frac{IFC * {}_kp_x * {}_kr_x}{(1+i)^{âge\ retraite\ -\ âge\ actuel}} \tag{1.4}$$

Celle-ci dépend de plusieurs facteurs détaillés ci-dessous :

- De l'**âge actuel** du salarié et de son **âge au moment de la retraite**. Le nombre d'années jusqu'à la retraite (***k***) est donc égal à :

$$k = âge\ retraite\ -\ âge\ actuel = N - n \tag{1.5}$$

Où N : l'année à la retraite.

- Des **droits** de l'employé au moment de la retraite. Ces droits sont généralement définies dans les conventions collectives nationales.

- Du **taux de charges** sociales (***g***, en pourcentages).

- Du **taux de revalorisation des salaires** qui est constant (***s***) et/ou dépend généralement de l'âge et de la catégorie professionnelle.

- Du **taux d'actualisation** (***i***).

- Du **salaire final** au moment de la retraite ($S_N$) :

$$S_N = S_n * (1+s)^{\hat{a}ge\ retraite\ -\ \hat{a}ge\ actuel} \tag{1.6}$$

Où $S_n$ : la salaire à la date d'évaluation.

- Du **montant de l'indemnité** versé à la retraite (***IFC***), qui est calculé de la façon suivante :

$$IFC = \frac{S_N}{12} * Droits * (1+g) \tag{1.7}$$

- De la **probabilité de survie** de l'individu d'ici à la retraite :

Celle-ci se calcule en fonction du sexe du salarié, de son âge actuel et de son âge à la retraite, à partir d'une table de mortalité publiée par l'INSEE et définie par les hypothèses de début de mission.

En effet, une table de mortalité présente, pour chaque âge x quelle contient :

- Un nombre d'individus vivants, éventuellement par sexe, par catégorie socioprofessionnelle, etc. Par convention noté $l_x$
- Une probabilité de décès dans l'année, par convention noté $q_x$

La probabilité d'un individu d'âge x de survive jusqu'à l'âge x+k est alors égale à :

$$_kp_x = \frac{l_{x+k}}{l_x} \tag{1.8}$$

- De la **probabilité de présence** dans l'entreprise à l'âge x+k :

En notant $r_x$ le nombre d'individu présents à un âge x donné par la table de turnover, éventuellement par catégorie sociaux-professionnelles, la probabilité que l'employé soit dans l'entreprise à l'âge x+k est de :

$$_kr_x = \frac{r_{x+k}}{r_x} \tag{1.9}$$

La table de turnover présente, à chaque âge associé, la probabilité que l'individu quitte son emploi. Elle est généralement définie empiriquement sur des statistiques communiquées par l'entité sur ses effectifs et est mise à jour en moyenne tous les trois ans. Il est courant que la table soit définie par âge et catégorie socioprofessionnelle.

**Calcul de prestations attendues :**

Le montant des prestations attendues (Expected Benefit Payments - EBP) est défini comme la valeur probable des indemnités à verser au moment du départ à la retraite du salarié dont la formule de calcul est comme suivante :

$$EBP_N = IFC * {}_kp_x * {}_kr_x \tag{1.10}$$

**Calcul de coût des services et coût d'intérêt :**

Le coût des services rendus (***SC)*** et le coût d'intérêt (***IC***) de l'année suivant celle de l'évaluation sont calculés par les formules suivantes :

$$SC_{n+1} = \frac{DBO_n}{Ancienneté\ actuelle}(1+i) \tag{1.11}$$

$$IC_{n+1} = \left(DBO_n - 0,5 * EBP_{n+1}\right) * i \tag{1.12}$$

Ces montants seront reconnus dans la charge (ou la dotation) de l'exercice suivant.

**L'engagement et les éléments de la charge de l'entreprise :**

Les montants totaux des DBO, SC et IC d'une entreprise sont donc simplement la somme de tous les individus de cette entreprise éligibles aux droits du régime à évaluer.

$$DBO_{Totale} = \sum_{i=1}^{M} DBO_i \quad ; \quad SC_{Total} = \sum_{i=1}^{M} SC_i \quad ; \quad IC_{Total} = \sum_{i=1}^{M} DBO_i \tag{1.13}$$

D'où M est le nombre total des salariés octroyés aux droits du régime.

# Chapitre 2

# Motivations de l'étude et modèle de référence

Ce chapitre a pour objectif de présenter les motivations de l'étude, de définir un processus global de l'étude et puis détailler ses étapes clés dans les chapitres suivants. Il présente enfin les résultats de l'évaluation des IFC d'un modèle de référence.

## 2.1 Contexte et motivations

Le turnover est un des sujets les plus discutables dans de nombreux domaines de recherche, et notamment en actuariat. Cette section présentera le contexte et les motivations de l'étude de turnover, et puis résumera les étapes principales du processus de l'étude.

### 2.1.1 Exigences de la norme IAS 19 et des commissaires aux comptes

L'hypothèse de turnover, ou taux de rotation en français, vise à prendre en compte dans l'engagement la probabilité que le salarié soit encore présent dans l'entreprise au moment où il devrait percevoir sa prestation (i.e. dans notre cas à son départ à la retraite). Cette hypothèse ayant un impact particulièrement significatif sur le niveau de l'engagement des IFC, elle fait donc l'objet d'une attention particulière par l'actuaire et par les commissaires aux comptes. Selon la norme IAS 19, les paragraphes 75 et 76 précisent que les hypothèses actuarielles doivent être non-biaisées, mutuellement compatibles et les meilleures estimations des toutes les variables qui détermineront le coût ultime des avantages postérieurs à l'emploi de l'entreprise. Plus précisément, les hypothèses actuarielles sont dites « non-biaisées » si elles ne sont ni imprudentes ni excessivement conservatrices, et sont dites « mutuellement compatibles » si elles reflètent les relations économiques entre des facteurs tels que l'inflation, les taux d'augmentation des salaires et les taux d'actualisation.

L'hypothèse de turnover n'étant pas clairement définie dans la norme IAS 19, il y a eu des débats sur les motifs de sortie à inclure dans le taux de turnover. En 2018, la CNCC a rédigé

une note qui précise que seuls les motifs de démissions doivent être pris en compte dans le cadre de l'évaluation des engagements sociaux (cf. **Annexe B**). A cet effet, les motifs à exclure dans le turnover sont : les décès, les licenciements économiques, les licenciements individuels, les mutations, les invalidités et inaptitudes, les ruptures conventionnelles et les fins de périodes d'essai. La note de CNCC a indiqué également que si la table de turnover ne tient compte que les démissions, l'entreprise a la possibilité de de reconnaître en compte de résultat l'engagement des salariés sortis via des licenciements individuels ou des ruptures conventionnelles dans la mesure où ils ne seront plus anticipés via le taux de sortie.

Par conséquent, il convient pour chaque entreprise d'établir une table de turnover reflétant le plus finement possible la réalité des taux de ***démissions*** uniquement observées au sein de cette entreprise et permettant de représenter les comportements sociaux et économiques sur la durée restante de l'engagement.

### 2.1.2 Les pratiques marché et les motivations d'une nouvelle étude

En pratique, les tables utilisées pour cette hypothèse sont construites en déterminant le taux de sortie à chaque âge pour la population concernée. Ces taux de sortie sont déterminés à partir des données transmises par l'entreprise, en se basant sur les deux ou trois années précédentes. Cette méthode permet de construire une table de turnover en fonction de l'âge et/ou de la catégorie. Cependant, de nombreux écarts sont observés entre le turnover choisi dans les hypothèses et le turnover réel, donnant lieu à la constatation systématique d'écarts actuariels. D'ailleurs, plusieurs facteurs qui pourraient avoir des impacts significatifs sur le taux de démission d'un salarié n'ont pas été pris en compte, par exemple : l'ancienneté, le salaire, la catégorie, le sexe, le statut matrimonial, la zone géographique, etc. Ainsi, une approximation trop grossière des taux de sortie peut engendrer des erreurs majeures et l'entreprise risque donc de sous ou surestimer l'engagement.

Par ailleurs, dans un contexte économique difficile, où les taux d'actualisation déterminés grâce aux taux des obligations de première catégorie du secteur privé sont historiquement bas, les provisions liées aux engagements sociaux s'avèrent parfois très élevées et constituent ainsi un véritable enjeu financier pour l'entreprise. Affiner l'hypothèse de turnover constitue un moyen d'affiner au plus précis la provision et permet de limiter les écarts actuariels. Il est à noter que la norme IAS 19 a été révisée en juin 2012 et que la méthode du « corridor » qui permettait d'amortir les écarts actuariels sur la durée résiduelle de l'engagement n'est désormais plus autorisée. Les écarts actuariels impactent ainsi directement les capitaux propres, ce qui encourage les entreprises à limiter ces écarts.

Les tables de taux de rotation actuelles de l'entreprise de l'étude, qui sont en fonction de l'âge, de la catégorie et de l'entité, ont été construites en 2016 en se basant sur les données historiques de démissions de 2014 à 2016. Étant donné que les écarts d'expérience liés à cette hypothèse ont été importants sur les dernières clôtures, et la structure des taux de démissions en fonction de l'âge n'a pas significativement changé depuis 2017, les méthodes statistiques comme le lissage du turnover par âge pourraient être inefficaces dans la détermination de tables de turnover.

Par conséquent, ce mémoire a pour but de donner une nouvelle approche pour l'actuaire en déterminant les tables de turnover au niveau individuel avec l'aide des modèles d'apprentissage automatique. En se basant sur les données de démissions observées dans la dernière année et les données RH des salariés de l'entreprise, le modèle d'apprentissage automatique va estimer la probabilité de rotation propre aux caractéristiques de chaque salarié, et puis projeter une courbe de taux tenant compte des évolutions dans le futur de ces caractéristiques.

Heureusement pour les actuaires et les entreprises, grâce à la rapidité des développements de l'intelligence artificielle, de la baisse des prix du stockage et de la puissance de calculs, les capacités de machine learning sont devenues de plus en plus accessibles (Shmueli, Patel, & Bruce, 2010 ; Witten, Frank, Hall & Pal, 2016). D'ailleurs, le volume de données disponibles de l'entreprise choisie est considérablement vaste qui améliorera généralement la qualité de prévision des modèles d'apprentissage automatique.

**Le schéma ci-dessous résume les étapes principales de l'étude :**

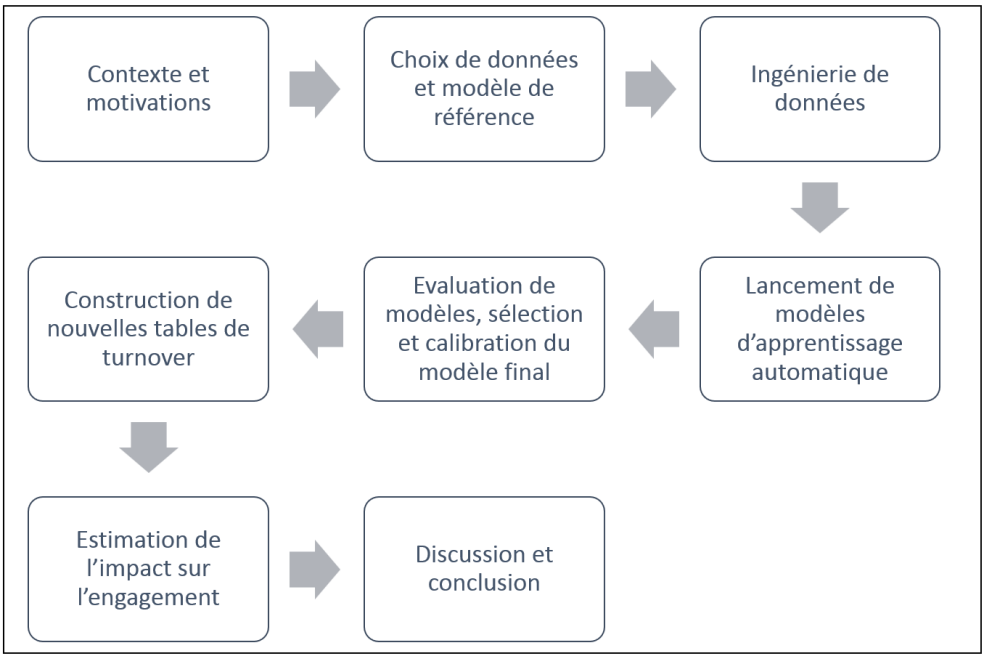



FIGURE 2.1 – Processus de l'étude

Le choix de données et les premiers résultats du modèle de référence seront présentés dans la sections suivante. Le chapitre 3 abordera les techniques de l'ingénierie des données et la préparation des données pour les modèles de prévision de taux de turnover. Le chapitre 4 introduira ensuite les algorithmes de machine learning utilisés pour prédire le turnover ainsi que les métriques d'évaluation de modèles. Enfin, le chapitre 5 présentera les résultats de l'étude y compris la construction de nouvelles tables de turnover, l'estimation de l'impact sur l'engagement et la conclusion.

## 2.2 Modèle de référence

L'objectif de cette section est de construire un modèle de référence qui sert à comparer les résultats obtenus par les nouvelles tables de turnover proposées dans cette étude. Cette section présente dans un premier temps le choix des données pour l'étude. Dans un second temps, elle introduit les hypothèses retenues pour le modèle de référence. Pour finir, elle résume les premiers résultats obtenus par ce modèle.

### 2.2.1 Le choix de données

Le but de ce mémoire est d'établir les tables de turnover au niveau individuel en utilisant les modèles d'apprentissage et d'étudier l'impact sur l'engagement d'une entreprise en particulier. Le choix des données est donc très important car les tables se basent par définition principalement sur les données et l'efficacité des algorithmes d'apprentissage automatique améliorera généralement avec la suffisance des données. En effet, trouver un client pour lequel une base suffisante de données soit disponible n'a pas été facile compte tenu du volume important de données à exploiter.

Pour rappel, selon la norme IAS 19, la constatation d'une provision de l'engagement des IFC est obligatoire et les sociétés cotées en Bourse dans l'Union européenne doivent obligatoirement présenter leurs comptes consolidés en normes IFRS. Compte tenu du cadre légal (les normes IAS 19 / IFRS), le Conseil National de la Comptabilité préconise la comptabilisation au bilan des engagements sociaux. Toutefois, que l'entreprise provisionne ou mentionne simplement ses engagements en annexe, elle doit évaluer les engagements. Ainsi, l'entreprise qui a choisi de ne pas provisionner ses engagements de retraite et engagements sociaux doit fournir en annexe la même qualité d'information établie sur des bases identiques à celle exigée des entreprises qui ont choisi de les provisionner.

Une entreprise choisie doit provisionner et comptabiliser ses Indemnités de Fin de Carrière en norme IAS 19. Aon compte un grand nombre de ces clients. Cependant, comme évoqué ci-dessus, il n'a pas été facile de trouver un client avec une base de données qui pourrait répondre aux besoins de l'étude.

Par coïncidence, lors de la clôture 2020, Aon a accompagné une entreprise française dont l'effectif est d'environ 23 mille salariés à revoir l'hypothèse de turnover qui sert aux calculs de son engagement des IFC. Conformément au règlement général sur la protection des données (RGPD) entré en vigueur le 25 mai 2018 dans toute l'Union européenne et la confidentialité convenue dans le contrat entre Aon et le client, plusieurs données ont été anonymisées et les autres informations comme le secteur d'activité ou la convention collective nationale de l'entreprise ne sont pas présentés dans ce mémoire.

L'étude est réalisée sur la population des salariés en contrat à durée indéterminée de cette entreprise. La base des données qui sert aux calculs des IFC a été construite à partir des données individuelles nécessaires (l'âge, l'ancienneté, le salaire, la catégorie socio-professionnelle, le sexe). Néanmoins, la base des données pour l'étude de turnover tient en compte de toutes les autres informations disponibles (notamment la société, la région, le statut de mariage, etc.) qui seront analysées en détail dans le chapitre 3.

Le tableau ci-dessous présente les principales caractéristiques de la population au 31/12/2019 :

| | **Statistiques au 31/12/2019** |
|---|---|
| **Effectif** | 23 298 |
| **Masse salariale** | 1 559 851 K€ |
| **Salaire moyen** | 66,952 K€ |
| **Age moyen** | 46,22 |
| **Ancienneté moyenne** | 17,44 |

Tableau 2.1 – Données statistiques de l'entreprise

Il est à noter que le salaire retenu pour les calculs des IFC est le montant le plus favorable entre le salaire annuel théorique et le salaire annuel reconstitué.

Avant de lancer les calculs actuariels, les tests de cohérence ont été réalisés sur les données reçues afin de s'assurer que celles-ci soient « valides ».

### 2.2.2 Les hypothèses retenues

Le choix des hypothèses actuarielles est crucial pour le calcul de l'engagement. En effet, certaines hypothèses peuvent avoir un impact significatif sur l'engagement comme le taux d'actualisation, et la table de turnover. Les différentes hypothèses retenues pour le modèle de référence sont présentées comme suit.

#### 2.2.2.1 Les hypothèses de calcul

La date d'évaluation est le 31/12/2019. Par ailleurs, l'inflation est fixée à 1,50 %. Le taux d'inflation utilisé est un taux à long terme et sa détermination est guidée par les publications de la Banque centrale européenne (BCE).

L'inflation permet de prendre en compte l'effet de l'érosion monétaire dans le temps. Dans les évaluations, le taux d'inflation est inclus dans les hypothèses telles que le taux d'actualisation, le taux de revalorisation des salaires et toute autre hypothèse de dérive monétaire.

#### 2.2.2.2 Les hypothèses financières

**Le taux d'actualisation :**

Le taux d'actualisation est le taux appliqué pour déterminer la valeur actuelle des prestations futures, il a donc un rôle clé dans les évaluations et doit être choisi avec précaution et en accord avec les normes comptables.

Le taux d'actualisation est défini dans la norme IAS 19 (§83 à 86) comme étant égal au taux des obligations émises par les entreprises de première catégorie à une échéance égale à la maturité de l'engagement. Dans le cas où il n'existe pas de marché pour ce type d'obligations, il faut se référer au taux de rendements du marché des obligations d'État (OAT) correspondantes.

Pour des groupes multinationaux, il faut tenir compte de taux d'actualisation différents suivant la zone monétaire et le pays dans lequel la société se trouve. Il est toutefois à noter que pour la zone Euro, les taux d'actualisation doivent être identiques ceci à cause de la convergence des politiques monétaires et budgétaires des États membres.

Il existe plusieurs méthodes pour déterminer le taux d'actualisation. Les deux les plus utilisées sont les suivantes :

- Un référentiel d'indice de taux dépendant de la duration du plan (les taux BLOOMBERG, IBOXX, OAT) ;
- Une courbe de taux des obligations à taux fixes sur la base des DBO futures estimées du plan.

Voici les graphiques des taux utilisés pour définir le taux d'actualisation pour les évaluations en norme IAS 19 :

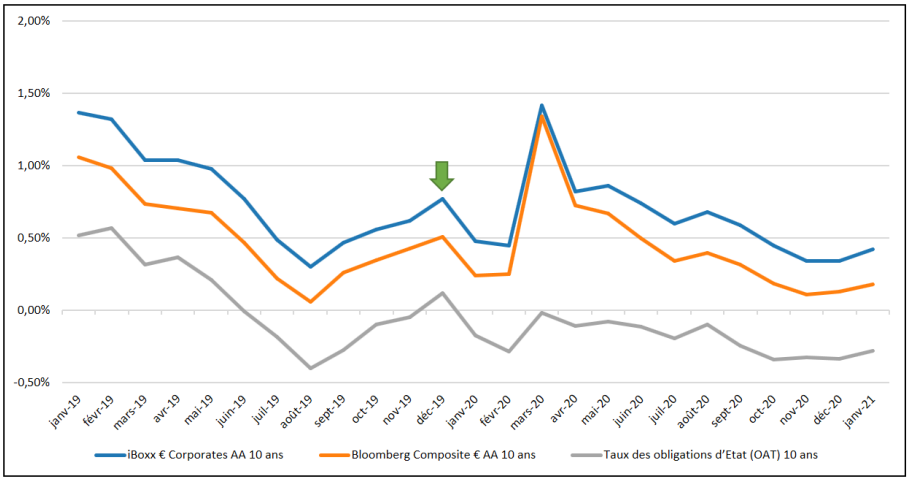



FIGURE 2.2 – Indice des taux à maturité 10 ans

Le taux utilisé pour l'évaluation au 31/12/2019 est de 0,75%.

**Le taux revalorisation des salaires :**

Dans l'évaluation de l'engagement des IFC, le taux de revalorisation des salaires sert à estimer le salaire à la retraite qui est utilisé pour calculer les droits des IFC.

Le paragraphe §90 dans la norme IAS 19 précise que les estimations de taux de revalorisation de salaires futures doivent prendre en compte l'inflation, l'ancienneté, les promotions et d'autres facteurs jugés pertinents. Il faut absolument distinguer les augmentations à long terme (à utiliser dans les évaluations actuarielles) et les augmentations à court terme retenues pour la détermination des budgets par exemple.

Dans cette étude, le taux de revalorisation de salaires est fixé à 2,50% pour les cadres et à 2,00% pour les non cadres.

**Le taux de charges sociales :**

Selon leur nature (départ volontaire, mise à la retraite, etc.), les Indemnités de Fin de Carrière sont soumises à des charges patronales qui font partie intégrante de l'engagement de

l'entreprise. Les charges sociales sont versées par l'entreprise lors du paiement de l'IFC du salarié. Les charges sociales sont des données communiquées par les services RH. Les taux de charges sociales varient entre 30% et 60% et souvent sont en fonction de la catégorie socio professionnelle.

Dans cette étude, le taux de charges sociales retenu est de 49,50%.

#### 2.2.2.3 Les hypothèses démographiques

**L'âge de départ à la retraite :**

L'âge de départ à la retraite correspond à l'âge du salarié lorsqu'il quitte l'entreprise pour liquider sa retraite. A ce moment, une indemnité de fin de carrière lui est versée par l'entreprise. L'âge légal de départ à la retraite des salariés en France dépend de plusieurs critères, et notamment de leur date de naissance. Cet âge minimum de départ est avancé pour les carrières longues ou les métiers pénibles, ou encore les salariés atteints de handicap.

Cette hypothèse se présente de plusieurs façons :

- Données RH de dates départs à la retraite qui sont propres aux pratiques de l'entreprise ;
- Age de départ à la retraite en fonction des catégories socioprofessionnelles (exemple cadre 65 ans, non cadre 62 ans) ;
- Calcul de l'âge à taux plein de départ à la retraite pour chaque individu en fonction de l'année de naissance, et du nombre de trimestres à cotiser, avec une hypothèse d'âge de début de carrière et un âge minimum de départ à la retraite fixé à 62 ans.

Les hypothèses d'âge de début de carrière sont souvent :

- Cadres : 23 ans ou 24 ans ;
- Non cadres : 20 ans ou 21 ans.

Avec le minimum entre âge réel d'entrée dans la société et l'hypothèse d'âge de début de carrière.

Dans notre étude, nous utilisons l'âge de départ à la retraite à taux plein avec l'hypothèse d'âge de début en fonction de l'année de naissance et de la catégorie (cf. **Annexe C**).

**La table de mortalité :**

Pour le calcul des IFC, une table de mortalité est utilisée pour déterminer la probabilité du salarié d'être en vie au moment du départ à la retraite. Le choix de tables de mortalité est

très important, tout comme l'hypothèse de turnover car si ces hypothèses sont mal calibrées par rapport aux populations, l'entreprise risque de sous ou surestimer l'engagement. La table de mortalité est défini dans la norme IAS 19 (§81 à 82) comme étant la meilleure estimation de taux de mortalité de l'ensemble des personnes concernées par le plan. Le paragraphe 82 impose également la nécessité de prendre en compte les évolutions des tables de mortalité au travers des facteurs de longévité.

Les tables de mortalité prises en compte dans les évaluations actuarielles sont des tables nationales fournies par l'INSEE et se basent sur toute la population française. En général, on prend les tables INSEE (table statiques Homme / Femme) pendant la phase en activité du salarié et les tables TPH TGF 05 (tables générationnelles Homme / Femme par année de naissance) pendant la phase de retraite des salariés.

Ces tables de mortalité étant appliquées dans le calcul pour l'ensemble des salariés de l'entreprise, sans distinction entre les différents groupes géographiques ou sociaux professionnels, l'engagement peut être sous ou sur-provisionné, néanmoins on considère que le taux de mortalité tend à s'uniformiser parmi les multiples groupes de la population.

Dans cette étude, les tables utilisées sont les tables INSEE H/F 2014-2016 (cf. **Annexe D**).

**La table de turnover :**

Le turnover exprime la rotation du personnel au sein d'une entreprise. Dans le calcul de l'engagement des IFC, cette hypothèse sert à estimer la probabilité que le salarié soit encore présent dans les effectifs au moment de son départ à la retraite (i.e. au moment où il perçoit son indemnité de fin de carrière).

Généralement, la table de turnover est construite à l'aide d'une table d'expériences établie par l'actuaire sur la base des données individuelles fournies par l'entreprise. L'hypothèse du taux de turnover est à nouveau similaire à l'hypothèse de mortalité, car les deux sont prises en compte dans le calcul de l'engagement à travers un système tabulaire. Le turnover est déterminé par de nombreux facteurs, principalement : l'âge, l'ancienneté et la catégorie. Comme évoqué dans la section précédente, cette hypothèse a un impact important sur l'engagement des IFC et donc fait l'objet de notre étude.

Les tables de turnover actuelles utilisées pour le modèle de référence distinguent par entité et par catégorie socioprofessionnelle avec les taux plus élevés pour les cadres. Par ailleurs, ces tables ont été construites en 2016 basées sur les données historiques de démissions de 2014 à 2016. Elles sont affichées dans le tableau suivant :

| Entité | S1000 | | S1888 | | S106 ; S2471 | | S001 ; S364 | | FR23 | | S126 | | FR1B | | S410 | | S2480 | | S122 | | S1777 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Âge | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres | Cadres | Non cadres |
| 20 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 8,37 % | 2,25 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 21 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 8,37 % | 2,25 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 22 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 8,37 % | 2,25 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 23 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 8,37 % | 2,25 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 24 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 8,37 % | 2,25 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 25 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 6,28 % | 1,68 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 26 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 6,28 % | 1,68 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 27 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 6,28 % | 1,68 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 28 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 6,28 % | 1,68 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 29 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 5,08 % | 0,30 % | 2,38 % | 0,32 % | 2,38 % | 0,32 % | 4,20 % | 0,24 % | 4,20 % | 0,24 % | 1,26 % | 0,52 % | 6,28 % | 1,68 % | 0,64 % | 0,57 % | 0,55 % | 0,09 % |
| 30 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 3,05 % | 0,30 % | 1,73 % | 0,32 % | 1,73 % | 0,32 % | 2,96 % | 0,18 % | 2,96 % | 0,18 % | 1,26 % | 0,35 % | 3,14 % | 0,84 % | 0,64 % | 0,35 % | 0,55 % | 0,09 % |
| 31 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 3,05 % | 0,30 % | 1,73 % | 0,32 % | 1,73 % | 0,32 % | 2,96 % | 0,18 % | 2,96 % | 0,18 % | 1,26 % | 0,35 % | 3,14 % | 0,84 % | 0,64 % | 0,35 % | 0,55 % | 0,09 % |
| 32 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 3,05 % | 0,30 % | 1,73 % | 0,32 % | 1,73 % | 0,32 % | 2,96 % | 0,18 % | 2,96 % | 0,18 % | 1,26 % | 0,35 % | 3,14 % | 0,84 % | 0,64 % | 0,35 % | 0,55 % | 0,09 % |
| 33 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 3,05 % | 0,30 % | 1,73 % | 0,32 % | 1,73 % | 0,32 % | 2,96 % | 0,18 % | 2,96 % | 0,18 % | 1,26 % | 0,35 % | 3,14 % | 0,84 % | 0,64 % | 0,35 % | 0,55 % | 0,09 % |
| 34 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 3,05 % | 0,30 % | 1,73 % | 0,32 % | 1,73 % | 0,32 % | 2,96 % | 0,18 % | 2,96 % | 0,18 % | 1,26 % | 0,35 % | 3,14 % | 0,84 % | 0,64 % | 0,35 % | 0,55 % | 0,09 % |
| 35 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 1,02 % | 0,30 % | 1,09 % | 0,32 % | 1,09 % | 0,32 % | 1,72 % | 0,12 % | 1,72 % | 0,12 % | 1,26 % | 0,18 % | 2,09 % | 0,56 % | 0,64 % | 0,13 % | 0,55 % | 0,09 % |
| 36 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 1,02 % | 0,30 % | 1,09 % | 0,32 % | 1,09 % | 0,32 % | 1,72 % | 0,12 % | 1,72 % | 0,12 % | 1,26 % | 0,18 % | 2,09 % | 0,56 % | 0,64 % | 0,13 % | 0,55 % | 0,09 % |
| 37 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 1,02 % | 0,30 % | 1,09 % | 0,32 % | 1,09 % | 0,32 % | 1,72 % | 0,12 % | 1,72 % | 0,12 % | 1,26 % | 0,18 % | 2,09 % | 0,56 % | 0,64 % | 0,13 % | 0,55 % | 0,09 % |
| 38 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 1,02 % | 0,30 % | 1,09 % | 0,32 % | 1,09 % | 0,32 % | 1,72 % | 0,12 % | 1,72 % | 0,12 % | 1,26 % | 0,18 % | 2,09 % | 0,56 % | 0,64 % | 0,13 % | 0,55 % | 0,09 % |
| 39 | 0,55 % | 0,09 % | 0,55 % | 0,09 % | 1,02 % | 0,30 % | 1,09 % | 0,32 % | 1,09 % | 0,32 % | 1,72 % | 0,12 % | 1,72 % | 0,12 % | 1,26 % | 0,18 % | 2,09 % | 0,56 % | 0,64 % | 0,13 % | 0,55 % | 0,09 % |
| 40 | 0,52 % | 0,09 % | 0,52 % | 0,09 % | 0,68 % | 0,15 % | 0,77 % | 0,21 % | 0,77 % | 0,21 % | 1,08 % | 0,07 % | 1,08 % | 0,07 % | 1,03 % | 0,09 % | 1,05 % | 0,28 % | 0,50 % | 0,13 % | 0,52 % | 0,09 % |
| 41 | 0,52 % | 0,09 % | 0,52 % | 0,09 % | 0,68 % | 0,15 % | 0,77 % | 0,21 % | 0,77 % | 0,21 % | 1,08 % | 0,07 % | 1,08 % | 0,07 % | 1,03 % | 0,09 % | 1,05 % | 0,28 % | 0,50 % | 0,13 % | 0,52 % | 0,09 % |
| 42 | 0,52 % | 0,09 % | 0,52 % | 0,09 % | 0,68 % | 0,15 % | 0,77 % | 0,21 % | 0,77 % | 0,21 % | 1,08 % | 0,07 % | 1,08 % | 0,07 % | 1,03 % | 0,09 % | 1,05 % | 0,28 % | 0,50 % | 0,13 % | 0,52 % | 0,09 % |
| 43 | 0,52 % | 0,09 % | 0,52 % | 0,09 % | 0,68 % | 0,15 % | 0,77 % | 0,21 % | 0,77 % | 0,21 % | 1,08 % | 0,07 % | 1,08 % | 0,07 % | 1,03 % | 0,09 % | 1,05 % | 0,28 % | 0,50 % | 0,13 % | 0,52 % | 0,09 % |
| 44 | 0,52 % | 0,09 % | 0,52 % | 0,09 % | 0,68 % | 0,15 % | 0,77 % | 0,21 % | 0,77 % | 0,21 % | 1,08 % | 0,07 % | 1,08 % | 0,07 % | 1,03 % | 0,09 % | 1,05 % | 0,28 % | 0,50 % | 0,13 % | 0,52 % | 0,09 % |
| 45 | 0,49 % | 0,08 % | 0,49 % | 0,08 % | 0,34 % | 0,00 % | 0,45 % | 0,10 % | 0,45 % | 0,10 % | 0,44 % | 0,02 % | 0,44 % | 0,02 % | 0,80 % | 0,00 % | 0,52 % | 0,14 % | 0,37 % | 0,13 % | 0,49 % | 0,08 % |
| 46 | 0,49 % | 0,08 % | 0,49 % | 0,08 % | 0,34 % | 0,00 % | 0,45 % | 0,10 % | 0,45 % | 0,10 % | 0,44 % | 0,02 % | 0,44 % | 0,02 % | 0,80 % | 0,00 % | 0,52 % | 0,14 % | 0,37 % | 0,13 % | 0,49 % | 0,08 % |
| 47 | 0,49 % | 0,08 % | 0,49 % | 0,08 % | 0,34 % | 0,00 % | 0,45 % | 0,10 % | 0,45 % | 0,10 % | 0,44 % | 0,02 % | 0,44 % | 0,02 % | 0,80 % | 0,00 % | 0,52 % | 0,14 % | 0,37 % | 0,13 % | 0,49 % | 0,08 % |
| 48 | 0,49 % | 0,08 % | 0,49 % | 0,08 % | 0,34 % | 0,00 % | 0,45 % | 0,10 % | 0,45 % | 0,10 % | 0,44 % | 0,02 % | 0,44 % | 0,02 % | 0,80 % | 0,00 % | 0,52 % | 0,14 % | 0,37 % | 0,13 % | 0,49 % | 0,08 % |
| 49 | 0,49 % | 0,08 % | 0,49 % | 0,08 % | 0,34 % | 0,00 % | 0,45 % | 0,10 % | 0,45 % | 0,10 % | 0,44 % | 0,02 % | 0,44 % | 0,02 % | 0,80 % | 0,00 % | 0,52 % | 0,14 % | 0,37 % | 0,13 % | 0,49 % | 0,08 % |
| 50 | 0,44 % | 0,07 % | 0,44 % | 0,07 % | 0,17 % | 0,00 % | 0,30 % | 0,05 % | 0,30 % | 0,05 % | 0,22 % | 0,01 % | 0,22 % | 0,01 % | 0,40 % | 0,00 % | 0,00 % | 0,00 % | 0,20 % | 0,13 % | 0,44 % | 0,07 % |
| 51 | 0,44 % | 0,07 % | 0,44 % | 0,07 % | 0,17 % | 0,00 % | 0,30 % | 0,05 % | 0,30 % | 0,05 % | 0,22 % | 0,01 % | 0,22 % | 0,01 % | 0,40 % | 0,00 % | 0,00 % | 0,00 % | 0,20 % | 0,13 % | 0,44 % | 0,07 % |
| 52 | 0,44 % | 0,07 % | 0,44 % | 0,07 % | 0,17 % | 0,00 % | 0,30 % | 0,05 % | 0,30 % | 0,05 % | 0,22 % | 0,01 % | 0,22 % | 0,01 % | 0,40 % | 0,00 % | 0,00 % | 0,00 % | 0,20 % | 0,13 % | 0,44 % | 0,07 % |
| 53 | 0,44 % | 0,07 % | 0,44 % | 0,07 % | 0,17 % | 0,00 % | 0,30 % | 0,05 % | 0,30 % | 0,05 % | 0,22 % | 0,01 % | 0,22 % | 0,01 % | 0,40 % | 0,00 % | 0,00 % | 0,00 % | 0,20 % | 0,13 % | 0,44 % | 0,07 % |
| 54 | 0,44 % | 0,07 % | 0,44 % | 0,07 % | 0,17 % | 0,00 % | 0,30 % | 0,05 % | 0,30 % | 0,05 % | 0,22 % | 0,01 % | 0,22 % | 0,01 % | 0,40 % | 0,00 % | 0,00 % | 0,00 % | 0,20 % | 0,13 % | 0,44 % | 0,07 % |
| >= 55 | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % | 0,00 % |

TABLEAU 2.2 – Les tables de turnover actuelles utilisées pour le modèle de référence

**La modalité de départ à la retraite :**

Dans cette étude, on suppose que tous les salariés partent volontairement à la retraite (initiative du salarié). Il est à noter que la mise à la retraite est interdite avant 70 ans et l'hypothèses d'âge de départ à la retraite se situe entre 62 ans à 67 ans, l'hypothèse de départ volontaire reste la plus cohérente.

#### 2.2.2.4 Les droits :

Les droits évalués correspondent à l'ensemble des prestations d'Indemnités de Fin de Carrière qui représentent un engagement pour la société vis-à-vis de ses salariés. Ces droits sont définis par les conventions collectives dont dépend l'entreprise. Dans cette étude, les droits définis dans la convention collective de l'entreprise sont exprimés en mois du salaire en fonction de l'ancienneté au moment du départ à la retraite de l'employé. Cependant, en raison de la confidentialité, les détails de la convention collective et les droits associés ne sont pas précisés dans cette étude.

Le tableau suivant résume les hypothèses présentées ci-dessus à retenir pour l'évaluation :

| **Hypothèses** | **2019** |
|---|---|
| Date de calcul | 31/12/2019 |
| Inflation | 1,50% |
| **Hypothèses financières** | |
| Taux d'actualisation | 0,75% |
| Taux de revalorisation des salaires (inflation incluse) | Cadres : 2,50%, Non cadres : 2,00% |
| Taux de charges sociales | 49,50% |
| **Hypothèses démographiques** | |
| Tables de mortalité | INSEE H/F 2014-2016 |
| Turnover | En fonction de l'âge et de la catégorie |
| Âge de début de carrière | En fonction de l'âge et de la catégorie |
| Âge de départ à la retraite | Âge taux plein |
| Modalité de départ | Départ volontaire |
| **Droits évalués** | En mois de salaire par ancienneté |

Tableau 2.3 – Les hypothèses actuariels retenues pour les évaluations au 31/12/2019

### 2.2.3 Premiers résultats

Chez Aon, le logiciel ProVal est utilisé pour calculer l'engagement des IFC. Ce logiciel s'adresse au personnel technique des compagnies de conseil pour calculer les engagements sociaux et simuler les modalités des régimes complémentaires de retraite. Néanmoins, ce logiciel n'est pas conçu pour calibrer les tables prospectives de turnover. Par conséquant, ProVal n'a pas été utilisé pour réaliser cette étude, et en fait, la modélisation de calculs des IFC a été réalisée sous l'aide d'Excel.

Le calcul de l'engagement de l'entreprise en appliquant les méthodes et les hypothèses présentées précédemment donne les résultats suivants (montants en milliers d'euros) :

| Catégorie | Effectif | Engagement DBO au 31/12/2019 | Service Cost 2020 | Interest Cost 2020 | Charge 2020 |
|---|---|---|---|---|---|
| Cadres | 9 837 | 567 510 | 34 746 | 4 223 | 38 969 |
| Non cadres | 13 461 | 392 196 | 21 235 | 2 890 | 24 125 |
| **Total** | **23 298** | **959 706** | **55 981** | **7 113** | **63 094** |

TABLEAU 2.4 – Engagement au titre des IFC calculé à partir des hypothèses 2019

Le montant que l'entreprise devra inscrire à son passif au titre des Indemnités de Fin de Carrière, c'est-à-dire la DBO, s'élève donc à **959 706** K€. La charge ou la dotation pour l'année 2020 s'élève à **63 094** K€ (Service Cost + Interest Cost).

Il est intéressant de découvrir la répartition de l'engagement de cette entreprise par tranche d'âge car cela permettra de comparer et d'expliquer les résultats obtenus par des nouvelles tables de turnover dans l'étude.

Le détail de l'engagement au 31/12/2019 par génération et pyramide des âges est présenté dans le graphique suivant :

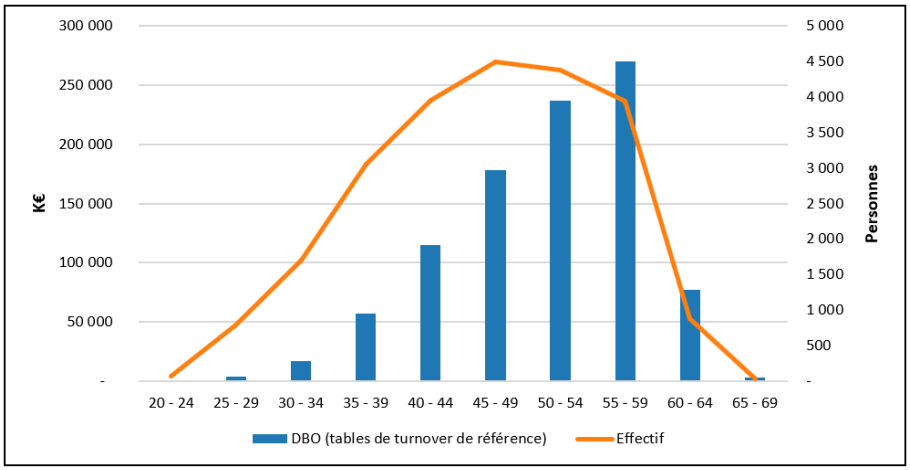



FIGURE 2.3 – Engagement au 31/12/2019 par tranche d'âges

Il est évident que l'engagement des salariés qui ont un âge entre 50 ans et 60 ans représente plus de la moité de l'engagement total car cette tranche d'âge présente les effectifs importants et un poids significatif sur l'engagement (horizon moyen court-terme pour un départ à la retraite).

Les prestations attendues dans les 10 prochaines années sont affichées dans le graphique

suivant :

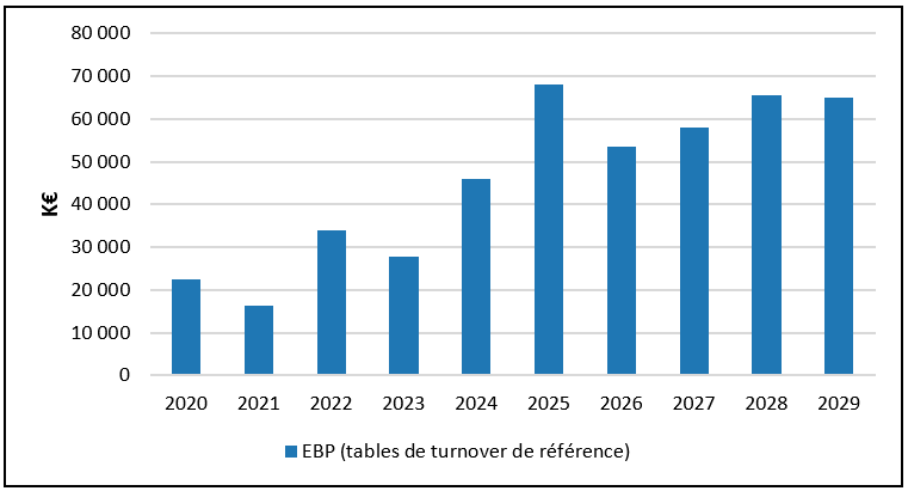



FIGURE 2.4 – Prestations futures attendues dans les 10 prochaines années

Le graphique montre que les prestations attendues augmentent significativement à partir de 2024.

Les résultats ci-dessus vont être utilisés comme les valeurs de référence pour comparer avec ceux obtenus par les nouvelles tables de turnover.

# Chapitre 3

# Ingénierie des données

L'objectif de ce chapitre est de présenter une étape essentielle de tout projet de marchine learning : l'ingénierie des données qui sert à mieux comprendre les caractéristiques de l'ensemble des variables des données, à lancer des techniques du traitement de données, appelées « le feature engineering » et enfin à améliorer les performances des algorithmes d'apprentissage automatique.

Dans un premier temps, ce chapitre fournira d'abord au lecteur une description et une analyse approfondies de l'ensemble des données utilisé dans cette étude. Il formera ensuite certaines remarques observées sur le turnover. Enfin, il présentera les techniques pour traiter et construire les bases des données.

## 3.1 Analyse des données

L'analyse de données fait référence au processus critique consistant à effectuer des enquêtes initiales sur les données afin de découvrir des tendances, de repérer des anomalies, de tester des hypothèses et de vérifier des hypothèses à l'aide de statistiques récapitulatives et de représentations graphiques.

L'analyse des données est principalement utilisée pour voir ce que les données peuvent révéler au-delà de la modélisation formelle et fournit une meilleure compréhension des variables de l'ensemble de données et des relations entre elles. Cela peut également aider à déterminer si les techniques statistiques utilisées l'analyse des données et les modèles d'apprentissage automatique sont appropriés.

### 3.1.1 Description des données

La base des données utilisée pour cette étude contient 23 298 observations, 22 variables explicatives et une variable à expliquer, à savoir si le turnover a eu lieu. Les variables explicatives comprennent des informations sur la démographie ainsi que certaines mesures financières. L'ensemble des données est exprimé en anglais et plusieurs informations ont été anonymisées.

Pour la bonne lecture, le tableau ci-dessous résume une description générale sur les variables présentes de la base des données :

| **Variables** | **Définition** | **Métrique** |
|---|---|---|
| Resignation | Si le salarié a démissionné | 1 : Démission, 0 : Non |
| EmployeeID | Identifiant unique pour l'individu | Numérique |
| Gender | Genre biologique | F (Femme) / M (Homme) |
| MaritalStatus | Statut matrimonial | 7 statuts |
| Category | Catégorie socioprofessionnelle | 15 catégories |
| ExeNonExe | Cadre ou Non cadre | Executive (Cadre) / Non Executive (Non cadre) |
| JobLevel | Niveau d'emploi | De 1 à 5 (5 indiquant le plus haut) |
| TypeOfEmployment | Type de contrat | 6 types de contrat |
| RTTRight | Droit à des Réduction du Temps de travail | Y (Oui) / N (Non) |
| Entity | L'entité pour laquelle le salarié travaille | 13 entités |
| City | La ville pour laquelle le salarié habite | 32 villes |
| Region | Le département pour lequel le salarié habite | 10 départements |
| Age | Nombre chronologique d'années qu'un individu a vécu | Numérique |
| YearsAtGroup | Ancienneté dans l'organisation actuelle | Numérique |
| TotalWorkingYears | Ancienneté carrière | Numérique |
| TheoreticalSalary | Salaire annuel théorique | Numérique |
| RestatedSalary | Salaire reconstitué hors primes exceptionnelles | Numérique |
| MonthlySalary | Salaire mensuel de base | Numérique |
| BonusRVI | Rémunération variable individuelle | Numérique |
| PercentRVI | Pourcentage de la RVI | Numérique |
| ExceptionalBonus | Prime exceptionelle | Numérique |
| NumberOfSalary-Months | Nombre de mois de salaire sur l'année | Numérique |
| PercentRemuneration | Pourcentage de la rémunération de base | Numérique |

Tableau 3.1 – Description des variables étudiées

### 3.1.2 Contrôle de validation des données

Pour augmenter la validité externe de cette étude, nous avons d'abord effectué quelques tests de base. Les six contrôles de base qui ont été testés sont :

- Des données manquantes ou vides ;
- $Age >= YearsAtGroup$ ;
- $Age >= TotalWorkingYears$ ;
- $TotalWorkingYears >= YearsAtGroup$ ;
- $TheoreticalSalary >= MonthlySalary$ ;
- L'ensemble des valeurs de salaires, de primes et de pourcentages devrait être non négatif.

Aucune indication de faible validité des ensembles de données internes n'a été trouvée, à l'exception de deux salaires annuels reconstitués négatifs de deux salariés. Cela pourrait être expliqué par le fait que ces salariés sont en absence (le nombre de mois de paie est égal à 0) et qu'ils ont des éléments de contributions ou cotisations déduites du salaire. Cela pourrait être une erreur de la collection des données du client. En effet, ceux deux valeurs négatives ont été rétablies à 0 lors du processus du traitement dans la section suivante.

### 3.1.3 La variable à expliquer

Les données des démissions sont observées durant l'année 2020 à partir de la population de l'entreprise au 31/12/2019. Dans cette base des données, seulement 4.13 %, 963 sur 23 298, ont été étiquetés comme ayant quitté la société en démission. Le taux de turnover est donc très bas. Il est à noter que le taux de turnover moyen en 2020 reste stable en comparant avec les taux des démissions observés sur l'historique des trois années précédentes. Selon les données trasmises par l'entreprise, depuis 2017, le taux de démissions moyen de cette entreprise est relativement bas et se situe entre 3,50 % et 4,50 % chaque année.

Étant donné le taux de turnover moyen très bas, même la règle de base « il est prévu que chaque employé reste dans l'entreprise » se traduira par une exactitude attendue de 95,87 % (100,00 % - 4,13 %). Cela pourrait affecter négativement les performances de certains des modèles prédictifs. Cependant, certaines méthodes existent pour contrer cela, elles seront discutées dans la section 4.4 du chapitre 4.

Le histogramme de la variable de réponse *Resignation* est donné dans le graphique suivant :

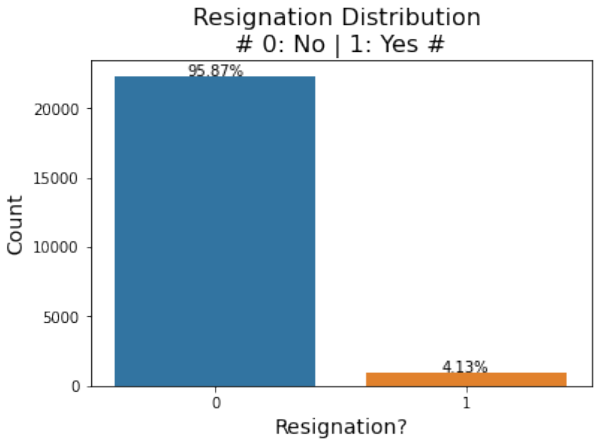



FIGURE 3.1 – Le taux de turnover historique

La variable à expliquer ayant seulement deux valeurs discrètes, les modèles de machine learning appropriés seront donc les modèles de classification qui sont présents en détail dans le chapitre 4.

Dans la base des données, il y a 22 variables explicatives, y compris 10 variables catégorielles et 12 variables numériques. L'ensemble des variables seront découvertes ci-dessous avec les analyses détaillées et ses relations avec la variable à expliquer, à savoir le turnover.

### 3.1.4 Les variables explicatives catégorielles

Les dix variables explicatives catégorielles sont : *Gender*, *MaritalStatus*, *Category*, *Exe-NonExe*, *JobLevel*, *TypeOfEmployement*, *RTTRight*, *Entity*, *City* et *Region*, dont la variable *JobLevel* est une variable ordonnée.

Tout d'abord, il est intéressant et important de vérifier la corrélation entre ces variables car cela permet de comprendre la relation entre les variables et pourrait puis suggérer des traitements nécessaires.

Afin de mesurer la corrélation entre les variables catégorielles, en statistiques, le test V de Cramér est le plus souvent utilisé. Il prend des valeurs comprises entre 0 et 1 (inclus), 0 correspondant à aucune association entre les variables et 1 correspondant à une variable étant complètement déterminée par l'autre.

Supposons qu'il y a deux variables catégorielles : $X$ prenant les modalités de $\{a_1, ..., a_I\}$ et $Y$ prenant les modalités de $\{b_1, ..., b_J\}$ et qu'il y a $n$ observations. Pour une combinaison

possible de valeurs $(a_i, b_j)$, soit $n_{i,j}$ = désignant nombre de fois où les valeurs $(a_i, b_j)$ ont été trouvées.

Les totaux des lignes et des colonnes sont donc :

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij}, \qquad n_{\bullet j} = \sum_{i=1}^{I} n_{ij} \tag{3.1}$$

La statistique du chi-carré est définie comme suit :

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{i,j} - n_{i\bullet} n_{\bullet j}/n\right)^2}{n_{i\bullet} n_{\bullet j}/n} \tag{3.2}$$

**La statistique V de Cramér**, communément désignée par $V$ mais parfois par $\phi_c$, est simplement une transformation de la statistique chi-carré :

$$V_{X,Y} = \sqrt{\frac{\chi^2/n}{min(I, J) - 1}} \tag{3.3}$$

La p-value pour la signification de V est la même que celle calculée à l'aide du test du chi-carré de Pearson.

Le tableau résumant les statistiques des corrélations V de Cramér entre les variables catégorielles dans cette étude est présenté en **Annexe E**. Pour la visualisation, les corrélations sont affichées dans le graphique suivant :

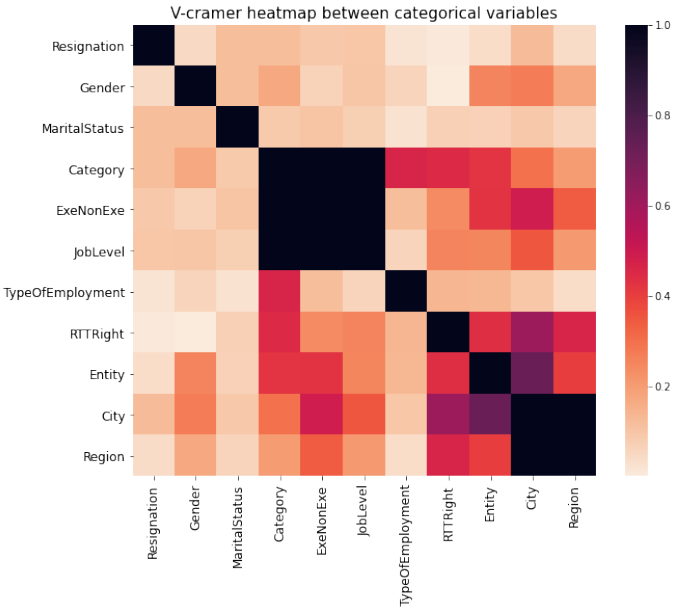



FIGURE 3.2 – Corrélations entre les variables catégorielles

Ce graphique mets en évidence le fait que les variables *Category*, *ExeNonExe* et *JobLevel* sont parfaitement corrélées, de même pour les variables la *City* et la *Region*. Quelques autres variables semblent fortement corrélées, par exemple *Category* et *RTTRight*, *Entity* et *City*.

Cette multicolinéarité causerait des problèmes potentiels car elle affaiblit la puissance statistique de certains modèles de régression ou de classification. Cependant, les variables corrélées pourraient encore avoir un impact sur la variable à expliquer et certains modèles d'apprentissage automatique serviraient à traiter la multicolinéarité.

Par ailleurs, l'analyse exploratoire de statistiques de chaque variable est essentiellement nécessaire car elle a pour but de mieux comprendre les données, d'examiner ses impacts sur la variable à expliquer et de donner au modeleur une idée initiale sur les possibilités du traitement de données.

Dans cette partie, les 2 variables catégorielles *MaritalStutus* et *TypeOfEmployment* seront découvertes pour fins d'illustrations.

***MaritalStutus***

Les statistiques de la variable *MaritalStutus* et le taux de turnover moyen par chaque statut marital peuvent être représentés par le graphique ci-dessous.

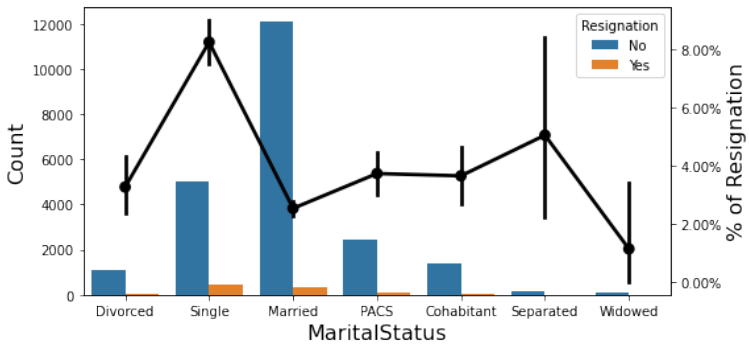



FIGURE 3.3 – Taux de turnover par *MaritalStatus*

Le graphique montre que :

- Les personnes qui habitent théoriquement « tout seules » (Divorcé, Célibataire, Séparé) ont une possibilité de démissionner plus élevés que ceux qui sont en couple (Marié, Pacsé, Concubiné) ;
- Les veufs/veuves et les personnes séparées ont très peu d'observations, 178 et 88 personnes respectivement. Il est donc nécessaire de regrouper ces salariés vers les autres groupes.

Les veufs/veuves pourraient être fusionnés vers les salariés mariés car le taux de turnover moyen des salariés mariés est le plus bas et inférieur au taux de turnover moyen de la population. Pour ceux qui sont séparés, il est logique de les regrouper avec la population de divorce.

***TypeOfEmployment***

Le graphique suivant présente la répartition de la population par type de contrat et le taux de turnover moyen associé.

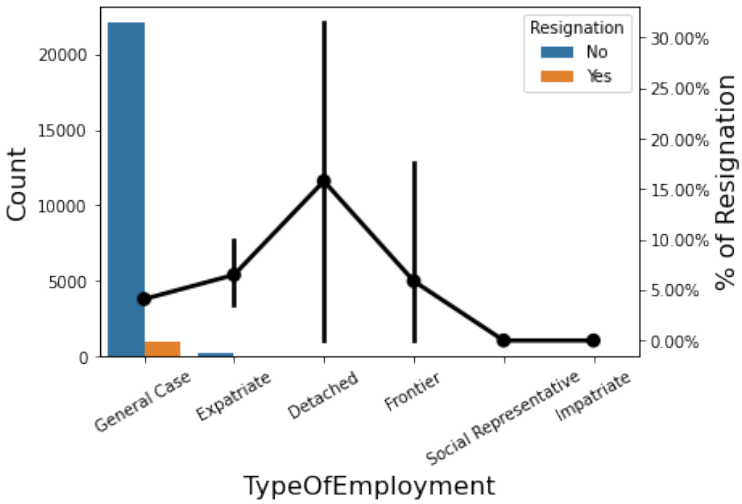



FIGURE 3.4 – Taux de turnover par *TypeOfEmployment*

Dans cette base de données, 98.99 % étant des contrats généraux (23 059 sur 23 298), cette variable est potentiellement retirée de l'étude car une variable qui est presque constante n'apporte rien d'essentiel dans l'étude statistique.

### 3.1.5 Les variables explicatives numériques

Dans cette base de données, il y a 12 variables explicatives numériques comme suivantes : *EmployeeID*, *Age*, *YearsAtGroup*, *TotalWorkingYears*, *TheoreticalSalary*, *RestatedSalary*, *MonthlySalary*, *PercentRVI*, *BonusRVI*, *ExceptionalBonus*, *NumberOfSalaryMonths*, et *PercentRemuneration*.

Lors de la première analyse, la variable *EmployeeID* s'est démarquée. Il est évident qu'il n'y a aucun lien de causalité entre *EmployeeID* et le turnover. Et comme les prédicteurs non

informatifs peuvent diminuer les performances de certains modèles (Kuhn & Johnson, 2016), cette variable est supprimée de la base des données.

Dans un premier temps, les principales statistiques de l'ensemble des variables explicatives numériques listées ci-dessus ont été observées et résumées par le tableau suivant :

| **Variables** | **Moyenne** | **Ecart type** | **Min** | **Q(25 %)** | **Q(50 %)** | **Q(75 %)** | **Max** |
|---|---|---|---|---|---|---|---|
| Age | 46,22 | 8,80 | 22,00 | 40,00 | 47,00 | 53,00 | 69,00 |
| YearsAtGroup | 17,44 | 9,70 | 1,00 | 11,00 | 17,00 | 25,00 | 46,00 |
| TotalWorkingYears | 23,60 | 9,32 | 1,00 | 17,00 | 24,00 | 31,00 | 48,00 |
| TheoreticalSalary | 58 700,64 | 33 661,32 | 0,00 | 41 137,80 | 50 945,58 | 64 957,38 | 1,20e+06 |
| RestatedSalary | 64 458,22 | 67 883,46 | -15 045,57 | 41 289,29 | 52 248,49 | 71 300,96 | 4,25e+06 |
| MonthlySalary | 4 400,08 | 2 965,57 | 0,00 | 2 667,33 | 3 530,95 | 5 230,56 | 1,00e+05 |
| PercentRVI | 4,71 | 7,88 | 0,00 | 0,00 | 0,00 | 10,00 | 150,00 |
| BonusRVI | 4 967,11 | 21 006,29 | 0,00 | 0,00 | 0,00 | 5 855,81 | 1,95e+06 |
| ExceptionalBonus | 1 613,07 | 22 874,43 | 0,00 | 0,00 | 0,00 | 0,00 | 2,66e+06 |
| NumberOfSalaryMonths | 11,05 | 2,44 | 0,00 | 12,00 | 12,00 | 12,00 | 12,00 |
| PercentRemuneration | 97,82 | 7,90 | 20,00 | 100,00 | 100,00 | 100,00 | 100,00 |

Tableau 3.2 – Statistiques des variables explicatives numériques

Les statistiques ci-dessus montrent que l'âge de la population se situe entre 22 et 69 ans avec une moyenne de 46,22 ans. L'ancienneté groupe moyenne et l'ancienneté carrière moyenne sont de 17,44 et 23,60 ans respectivement.

Comme évoqué dans la sous-section 3.1.2 précédente, il y a quelques valeurs négatives de salaire reconstitué qui peuvent être appelées les valeurs aberrantes. Il faudrait donc traiter ces valeurs. Il ne faut pas les confondre avec les valeurs extrêmes qui reflètent la réalité. Dans cette base des données, il y a les valeurs très élevées, dites extrêmes, des éléments de salaries et de primes qui pourraient impacter significativement les résultats. Certes, les algorithmes complexes peuvent traiter des bases de données brutes, avec leurs erreurs et leurs valeurs extrêmes, dans un cadre de *machine learning*.

Il faudrait également faire attention à l'asymétrie des variables *PercentRVI*, *NumberOfSalaryMonth* et *PercentRemuneration* qui reflète aussi la réalité, mais peut réduire la performance des algorithmes d'apprentissage automatique et impacterait donc négativement sur les résultats des coefficients estimés.

Si les données comportent des anomalies ou se trouvent l'asymétrie, les conclusions de l'étude risquent d'être biaisées où l'importance de traiter les données en amont. Par conséquent, quelques méthodes du traitement des variables numériques ont été réalisées et présentées dans la sous-section 3.2.3 suivante.

Dans un second temps, l'analyse des relations entre les variables numériques a été réalisée. La corrélation linéaire Pearson est utilisée dans cette étude. Pour rappel, le coefficient de corrélation de Pearson d'une paire de variables aléatoires $(X, Y)$ est généralement représenté par la lettre grecque $\rho$, et est défini par la formule ci-dessous :

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.4}$$

où :

- $cov$ la covariance ;
- $\sigma_X$ et $\sigma_Y$ l'écart type de X et de Y respectivement.

Lors de l'application à un échantillon de n observations constituées de n paires de données $\{(x_1, y_1), ..., (x_n, y_n)\}$, le coefficient de corrélation de Pearson est calculé comme suit :

$$r = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^n (x_i - \overline{x})^2}\sqrt{\sum_i^n (y_i - \overline{y})^2}} \tag{3.5}$$

avec $\overline{x} = \frac{1}{n}\sum_i^n x_i$.

Les corrélations parmi les variables explicatives numériques et le turnover sont présentées dans le graphique suivant :

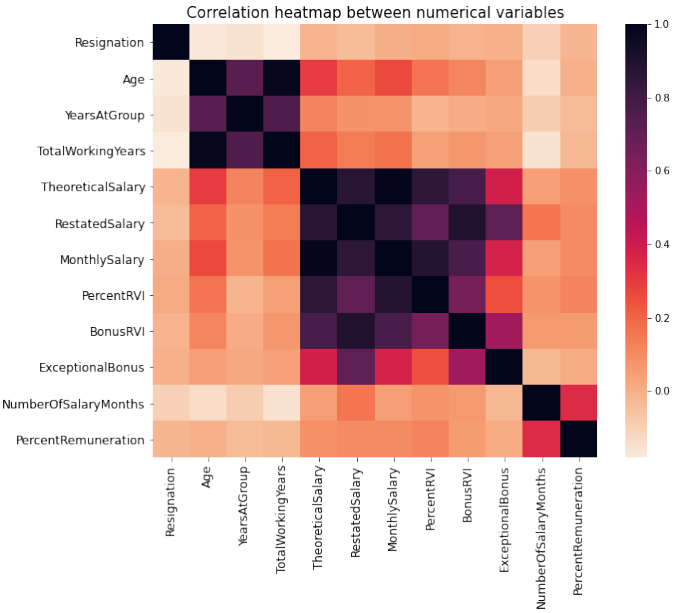



FIGURE 3.5 – Corrélations entre les variables numérique

Le tableau résumant les statistiques des corrélations Pearson entre les variables numérique est présenté en **Annexe E**. Il est évident et logique que les variables de rémunération et de primes sont fortement corrélées. Un salarié ayant un salaire mensuel de base élevé aura un salaire annuel théorique, reconstitué et les montants des primes élevés.

De même, *Age*, *YearsAtGroup* et *TotalWorkingYears* sont fortement corrélées. Logiquement, l'ancienneté carrière est positivement relative à l'âge du salarié. Le graphique suivant donne une analyse plus approfondie sur la corrélation de ces 3 variables en regardant leurs distributions par le turnover.

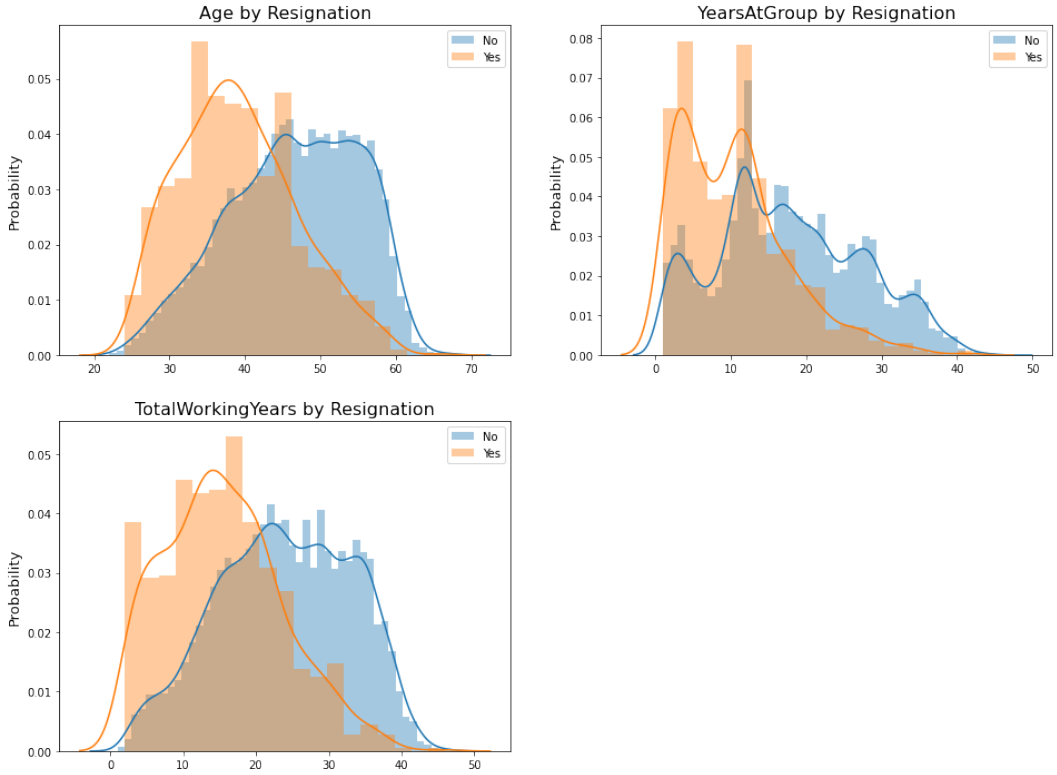



FIGURE 3.6 – Distribution par le turnover des variables *Age*, *YearsAtGroup* et *TotalWorkingYears*

Graphiquement, la variable *Age* a un impact important sur le turnover et a également une corrélation forte ($r = 0,9840$) avec la variable *TotalWorkingYears* qui renforce l'argument précédent.

Il est important de noter que l'efficacité des changements ou des traitements de données sur la performance de modèles d'apprentissage automatique devrait être mesurée par rapport avec la performance sur la base initiale de données. Dans la sous-section 3.2.2 suivante, quelques modèles de base seront réalisés afin de vérifier l'importance de chaque variable et chaque modalité des variables et puis proposer des traitements appropriés.

### 3.1.6 Remarques

D'après la combinaison des analyses ci-dessus et les figures dans l'**Annexe F**, nous pouvons souligner les points exploratoires comme suit :

**Remarque 1 : Le sexe est un prédicteur assez significatif du turnover.** Le taux moyen de turnover des femmes est de 5,43 % qui est plus élevé que celui des hommes, soit 3,25 %. Lorsque $V = 0,05$, le sexe est donc un impact assez important du turnover. Cela peut s'expliquer par le fait que les hommes du secteur d'activité de l'entreprise sont moins susceptibles de démissionner volontairement, bien que leur impact sur le turnover ne soit que modéré.

**Remarque 2 : La catégorie est un prédicteur important du turnover.** Le taux moyen de turnover des cadres est de 6,75 % qui est significativement élevé que celui des non cadres, soit 2,60 %. Il est évident que les cadres sont plus susceptibles de démissionner que les non cadres car le marché de l'emploi est très porteur pour les managers, ce qui leur donne l'opportunité de trouver un meilleur poste ailleurs. Lorsque $V_{Resignation,ExeNonExe} = 0,09$, cela induit à l'importance de la variable *Category* sur la prévision de turnover.

**Remarque 3 : La ville est un prédicteur important du turnover.** Au sein de l'entreprise, le turnover est significativement impacté par la ville d'habitation du salarié avec une corrélation $V_{Resignation,City}$ égale à 0,12. Les salariés dans les grandes villes sont plus susceptibles de démissionner que ceux qui habitent dans les plus petites villes.

**Remarque 4 : L'âge est négativement lié au turnover.** L'âge démontre l'effet le plus fort sur le turnover parmi les attributs individuels avec une corrélation $r_{Resignation,Age} = -0,17$. L'effet peut être interprété comme le fait que les salariés âgés sont moins susceptibles de démissionner.

**Remarque 5 : L'ancienneté groupe est négativement liée au turnover.** L'ancienneté groupe a un effet important sur le turnover avec un $r_{Resignation,YearsAtGroup} = -0,15$, ce qui peut être interprété comme le fait que les employés qui travaillent pour l'entreprise depuis plus longtemps sont moins susceptibles de démissionner.

**Remarque 6 : La rémunération est négativement liée au turnover.** Il est constaté que la rémunération affectait considérablement le turnover, ce qui indique que les employés qui reçoivent une rémunération plus élevée sont moins susceptibles de quitter volontairement l'entreprise. Dans l'étude, la rémunération se réfère aux salaires annuels (*TheoreticalSalary*, *RestatedSalary*) et aux primes (*BonusRVI*, *ExceptionalBonus*).

## 3.2 Traitement des données

Les traitements de données représentent une partie importante dans un projet lié au *Data Science* car ils permettent d'améliorer la performance des modèles de machine learning et donc de fournir de meilleurs résultats.

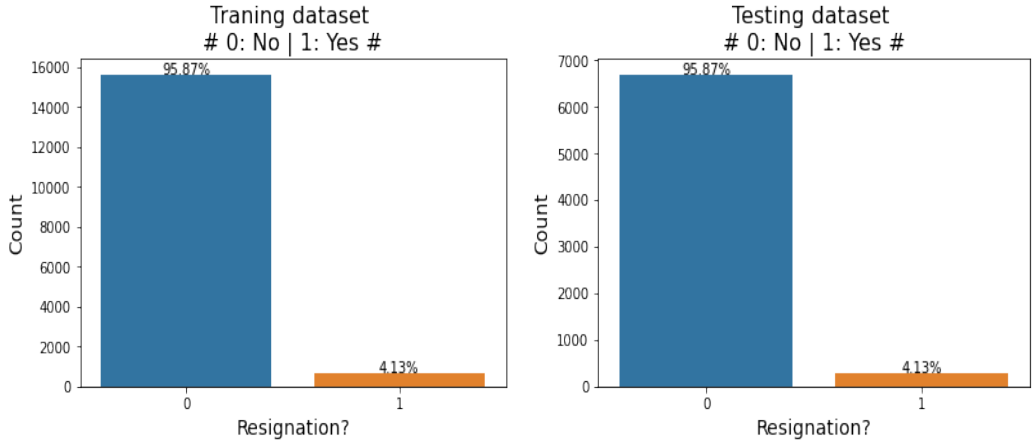
Les étapes du traitement des données peuvent s'articuler comme suit :

- Division des données. On divise d'abord l'ensemble des données en deux sous-échantillons : (i) échantillon d'apprentissage destiné à l'apprentissage des modèles et (ii) échantillon de test destiné à évaluer des modèles. Notons aussi que tous les traitements décrits dans la suite doivent être effectués de la même manière sur les deux sous-échantillons. Par exemple, pour normaliser une variable $X$, sa moyenne empirique et son écart-type empirique seront calculés sur l'échantillon d'apprentissage, puis seront également utilisés pour transformer observations de $X$ dans l'échantillon de test.
- Modèles naïfs. Avant de procéder aux traitements des données, nous exécutons l'ensemble des modèles naïfs sur l'échantillon d'apprentissage original en vue d'avoir un indice de référence.
- Traitement des données. On procède aux traitements des données en nous basant sur les analyses et les remarques de la section précédente.

### 3.2.1 Préparation de jeu des données

Il est inapproprié de valider les modèles sur les mêmes données avec lesquelles le modèle est entraîné, puisque les données ne sont pas nouvelles pour le modèle de prévision, les résultats seront biaisés et montreront une performance trop optimiste du modèle. Ainsi, pour valider correctement les modèles de prévision entraînés, il est nécessaire d'avoir une base de test (c'est-à-dire un échantillon que le modèle de prévision n'a pas vu auparavant). Nous avons choisi de diviser les données en une base d'apprentissage et une base de test de 70 % et 30 % respectivement. Le taux de turnover étant sous-représenté, un échantillonnage aléatoire stratifié est utilisé pour diviser les données. Cela garantira que les proportions de turnover sont égales dans chaque base des données. Le nombre d'enregistreurs pour la base d'apprentissage et la base de test est de 16 308 et 6 990 respectivement.

Le graphique suivant montre que les taux de turnover pour les 2 échantillons des données d'apprentissage et de test sont égaux et identiques au taux de turnover moyen de toute la population :




(A) Base des données d'apprentissage (B) Base des données de test

FIGURE 3.7 – Répartition des bases données par le turnover

### 3.2.2 Modèles naïfs

Lors de l'analyse des données, plusieurs « problèmes » sur la base des données initiale ont été détectés comme les suivants : la multicolinéarité et la corrélation significativement fortes parmi les variables, les modalités de certaines variables catégorielle avec peu d'observations, les valeurs aberrantes, extrêmes, etc.

Il y avait quelques idées pour résoudre ces problèmes comme le regroupement / la fusion des modalités d'une variable catégorielle, le retrait de certaines variables ou d'autres traitements sur les variables numériques. Cependant, le fait d'appliquer ces traitements basés seulement sur les statistiques observés semble être biaisé et manqué de preuves. Afin de vérifier l'efficacité et la pertinence des traitements de données, la technique de lancer quelques modèles naïfs est fréquemment utilisée en pratique.

Un modèle naïf, également appelé « modèle de base » est un modèle simple appliqué sur une base de données la plus moins manipulée (i.e. les données initiales), dont ses résultats ont pour but d'analyser l'importance des prédicteurs sur la variable à expliquer et de fournir une référence de comparaison avec les résultats des autres algorithmes de machine learning.

Les trois modèles naïfs qui ont été utilisés dans cette étude sont les suivants : Logistique non-pénalisée, CART (L'abre de décision) et K-Nearest Neighbors (Les K plus proches voisins - KNN). La régression logistique est une méthode standard pour la classification dû à sa simplicité et son efficacité. Concernant le modèle des arbres de décision, hormis sa simplicité et sa capacité d'interprétation, un avantage intéressant de ce modèle est qu'il est non paramétrique et ne

nécessite pas que les données soient normalement distribuées. Une des raisons pour laquelle le modèle KNN a été choisi est qu'il n'y a pas non plus de calculs complexes sur l'échantillon d'apprentissage et qu'il ne nécessite pas des hypothèses pour la distribution des modalités. La présentation de ces modèles ainsi que la description des métriques d'évaluation sont reportées au chapitre suivant.

La métrique d'évaluation des modèles naïfs est donnée par le tableau suivant :

| Model | Accuracy | Sensitivity | Specificity | Precision | Balanced Accuracy | F1 score | AUC |
|---|---|---|---|---|---|---|---|
| Logistic | | | | | | | |
| Train | 0,9586 | 0,0015 | 0,9999 | 0,3333 | 0,5007 | 0,0030 | **0,7879** |
| Test | 0,9582 | 0,0000 | 0,9996 | 0,0000 | 0,4998 | 0,0000 | **0,8285** |
| CART | | | | | | | |
| Train | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | **1,0000** |
| Test | 0,9332 | 0,2388 | 0,9631 | 0,2184 | 0,6009 | 0,2281 | **0,6009** |
| K-Neighbors | | | | | | | |
| Train | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | **1,0000** |
| Test | 0,9552 | 0,0242 | 0,9954 | 0,1842 | 0,5098 | 0,0428 | **0,5922** |

TABLEAU 3.3 – Métrique d'évaluation des modèles naïfs

En regardant les valeurs de l'aire sous la courbe ROC (*l'AUC* en anglais), la performance du modèle Logistique non-pénalisée est acceptable. Le modèle d'arbre de décision (étiqueté CART) et le modèle de KNN se rencontrent le problème de sur-apprentissage. Les résultats de prévision de ces modèles semblent faibles. La raison pour laquelle la valeur AUC a été utilisée pour la comparaison des modèles sera expliquée dans la section 4.5 du chapitre 4. Les graphiques suivants montreront les courbes ROC par modèle et par jeu de données.

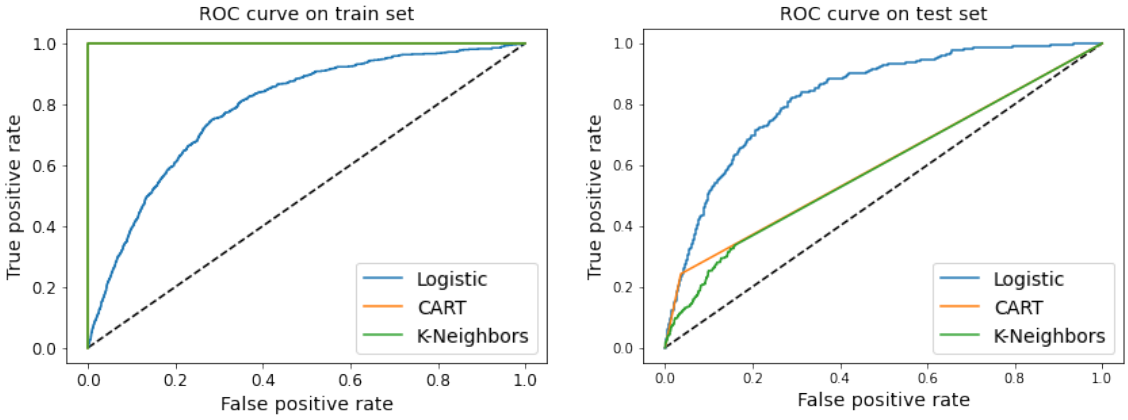



FIGURE 3.8 – ROC des modèles naïfs

Il est à noter que les résultats ci-dessus des modèles naïfs seront utilisés comme une référence de base pour évaluer la performance et l'amélioration des autres modèles de *machine learning*. Par exemple, si un autre modèle qui donne l'AUC plus élevée que celle du modèle Logistique non-pénalisé, cela signifie que la prévision de ce modèle est meilleur.

Pour l'objectif de vérifier l'importance de chaque variable et de chaque modalité des variables catégorielles et puis de proposer les traitements appropriés, les graphiques suivants donnent les meilleurs prédicteurs en faisant le *bagging* cent fois des modèles Logistique et CART.

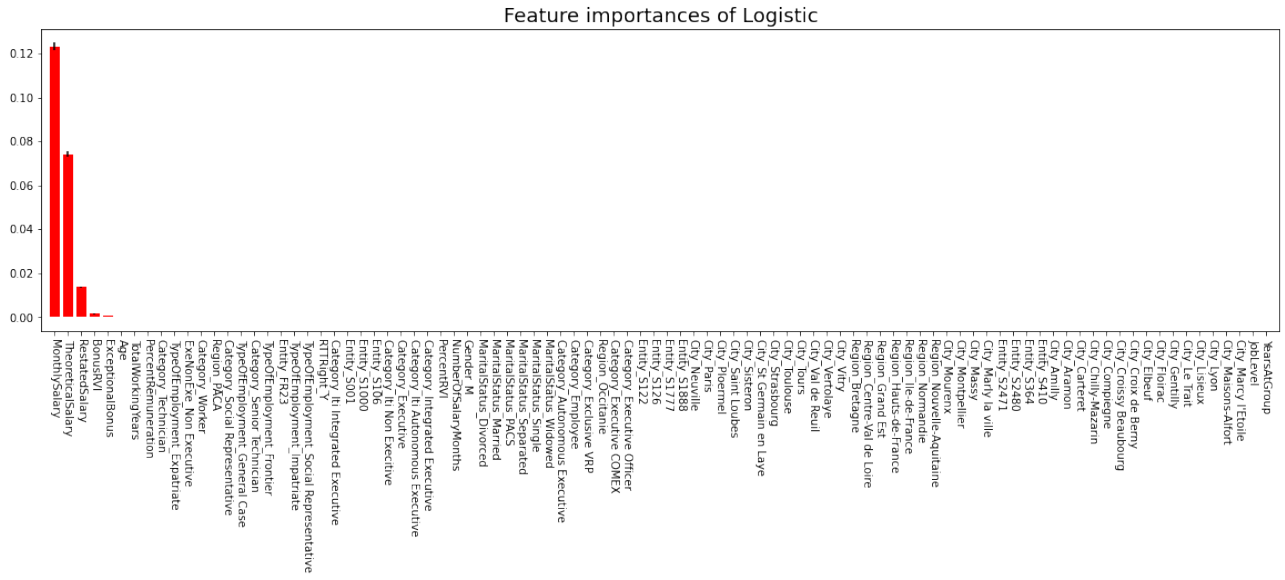



FIGURE 3.9 – Les meilleurs prédicteurs du modèle Logistique

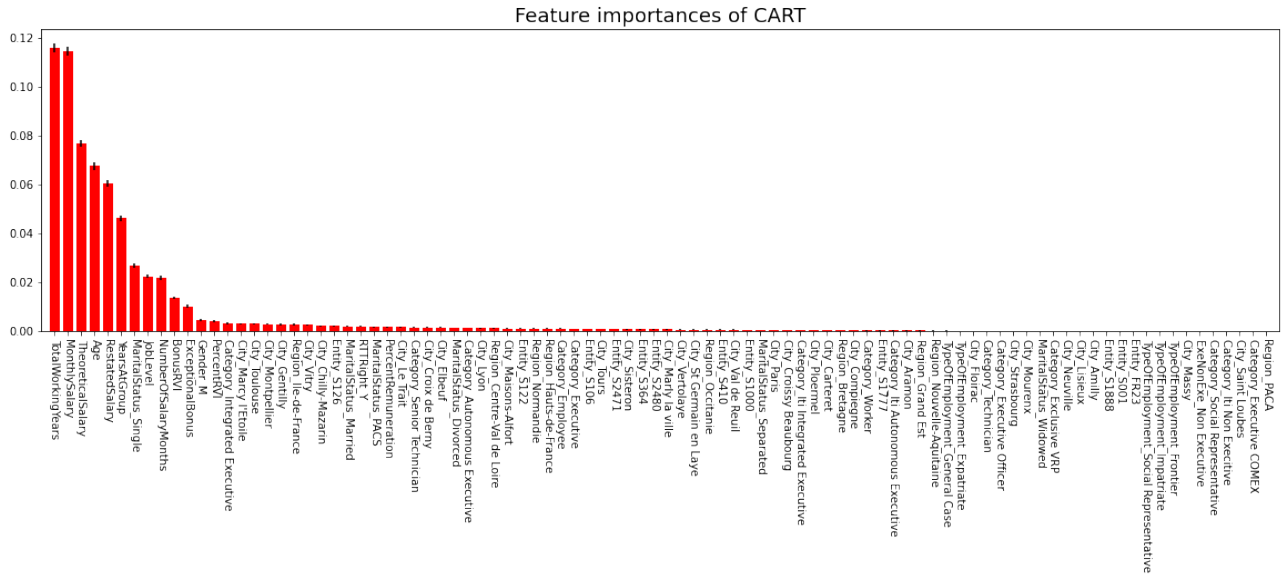



FIGURE 3.10 – Les meilleurs prédicteurs du modèle CART

Selon le modèle logistique, très peu de variables ont un impact sur la variable de réponse, tandis que le CART trouve plusieurs prédicteurs significatifs. Ce résultat est cohérent avec celui du tableau 3.3. Plus précisément, parce que le modèle logistique n'est pas capable d'exploiter la richesse d'information des données, il provoque donc le problème de sous-apprentissage ou *underfitting* (le métrique d'évaluation sur l'échantillon de test est meilleur que celui sur l'échantillon d'apprentissage). Alors que le CART capture bien l'information des données, mais il est beaucoup adapté au échantillon d'apprentissage, il s'agit donc d'un problème de sur-apprentissage ou *overfitting*. D'autre part, le graphique des variables importantes du modèle CART montre plus d'informations qui sont en ligne avec les analyses des données et remarques dans les sections précédentes. Grâce aux graphiquex si-dessus, il est convenu que :

- la variable *TypeOfEmployment* n'a pas d'impact sur le turnover et elle est donc retirée de la base des données ;
- les variables explicatives numériques ont un impact important sur le turnover ;
- quelques modalités de certaines variables explicatives catégorielles ont besoin de regrouper et fusionner.

### 3.2.3 Traitements des données

L'ensemble des techniques de traitements des données sera présenté comme suit :

#### 3.2.3.1 Traitement des variables explicatives catégorielles

Chaque variable explicative catégorielle va d'abord être regroupée si elle compose un grand nombre de modalité ou elle contient une modalité très peu fréquentée, puis va être transformée en numérique. Voir le détail comme suit.

**Regroupement / Fusion des modalités :**

Basé sur les analyses précédentes, il est évident de fusionner les modalités qui ont peu d'observations ou de regrouper les modalités qui devraient être statistiquement et logiquement liées. L'ensemble des traitements est proposé comme suit :

***MaritalStatus :***

- Fusionner les personnes *Widowed* vers le groupe des personnes *Maried* ;
- Fusionner les salariés avec statut *Separated* vers le statut *Divorced*.

***Category :***

- Fusionner les *Executive COMEX*, *Social Representative* et *Technician* vers le groupe des *AM* ;
- Regrouper les *Iti Autonomous Executive*, *Exclusive VRP*, *Employee* et *Iti Integrated Executive* dans un nouveau groupe *Other*.

***Entity :***

- Fusionner les salariés des entités *S001*, *1888*, *FR23* et *S1777* vers l'entité *S1000* ;
- Fusionner les autres entités avec moins de 250 observations (*S2480*, *S2471* et *FR1B*) en entité *S364*.

***City, Region :***

Les 2 variables étant parfaitement corrélées et ayant de nombreuses modalités, il est en effet nécessaire de créer une nouvelle variable *Zone* (*Zone = Region + City*) et de regrouper les modalités de cette nouvelle variable. Les détails du traitement sont présentés en **Annexe F**.

Il est important de noter que le traitement des modalités se base également sur l'expérience du modeleur. Après avoir lancé les modèles, le modeleur peut toujours faire un deuxième tour de traitement en fonction des résultats observés et puis vérifier l'efficacité des mises à jour. Dans cette étude, des autres possibilités des traitements de modalités ont été essayées, néanmoins les résultats ne sont pas significativement améliorés. En raison du compromis entre une amélioration marginale des résultats et le temps et les efforts consacrés, les traitements ci-dessus sont utilisés pour le modèle final.

**Numérisation :**

La plupart des algorithmes d'apprentissage automatique ne fonctionnent qu'avec des données numériques ; par conséquent, les variables binaires ont été créées. Au cours de ce processus, les catégories sont ré-encodées en petits bits d'information appelés « variables binaires ». Chaque catégorie a sa propre variable fictive qui est un indicateur zéro / un pour ce groupe. Une variable avec quatre modalités peut être transformée en seulement 3 nouvelles variables binaires, puisque la quatrième peut être déduite.

#### 3.2.3.2 Traitement des variables explicatives numériques

Le but de faire des traitements pour les variables explicatives numériques est d'améliorer la performance et la stabilité des calculs dans certains algorithmes d'apprentissage automatique. Les deux techniques de standardisation et de transformation Box-Cox sont présentées comme ci-après :

**Standardisation :**

La transformation de données la plus simple et la plus courante consiste à centrer et à mettre à l'échelle les variables de prédiction. Il s'agit de procédures de « preprocessing » effectuées pour certains modèles, appelées également « standardisation ». Pour centrer une variable de prédicteur, la valeur moyenne de prédicteur est soustraite de toutes les valeurs. Il en résulte que le prédicteur a une moyenne nulle. Pour mettre à l'échelle une variable de prédicteur, chaque valeur de la variable de prédicteur est divisée par son écart type. La mise à l'échelle des données oblige les valeurs à avoir un écart type commun de un. Ces manipulations sont généralement utilisées pour améliorer la stabilité numérique de certains calculs. Certains modèles bénéficient du fait que les prédicteurs sont à une échelle commune (Kuhn & Johnson, 2016). Le seul réel inconvénient de ces transformations est une perte d'interprétabilité des valeurs individuelles, puisque les données ne sont plus dans les unités d'origine.

Mathématiquement, supposons que la variable $X$ a $n$ observations dans un échantillon. La transformation « standardisation » de l'observation $x_i$, notée $z_i$, est définie comme la suivante :

$$z_i = \frac{x_i - \overline{x}}{\sigma} \tag{3.6}$$

où

- $\overline{x} = \frac{1}{n}\sum_i^n x_i$ ;
- $\sigma = \sqrt{\frac{1}{n}\sum_i^n (x_i - \overline{x})^2}$.

**Transformation Box-Cox :**

Une autre étape de « preprocessing » utilisée dans cette étude est une transformation proposée par (Box & Cox, 1964). Le but de cette transformation, dans cette étude, est de supprimer l'asymétrie distributionnelle dans les variables explicatives numériques. Une distribution non asymétrique est une distribution à peu près symétrique. Cela signifie que la probabilité de tomber de chaque côté de la moyenne de la distribution est à peu près égale. Une distribution asymétrique à droite a un grand nombre de points sur le côté gauche de la distribution (valeurs plus petites) que sur le côté droit (valeurs plus grandes).

La formule de la statistique d'asymétrie de l'échantillon est comme suit :

$$asymétrie = \frac{\sum_i^n (x_i - \overline{x})^3}{(n-1)v^{3/2}} \tag{3.7}$$

avec : $v = \frac{1}{n-1}\sum_i^n (x_i - \overline{x})^2$

où $X$ est la variable explicative, $n$ est le nombre d'observations et $\overline{x}$ est la moyenne de l'échantillon de la variable. Si la distribution du prédicteur est à peu près symétrique, les valeurs d'asymétrie seront proches de zéro. Alors que la distribution devient plus asymétrique à droite, la statistique d'asymétrie devient plus grande. De même, à mesure que la distribution devient plus asymétrique à gauche, la valeur devient négative.

Le remplacement des données par le log, la racine carrée ou l'inverse peut aider à supprimer l'asymétrie. Une méthode statistique utilisée pour identifier empiriquement une transformation appropriée est la méthode (Box & Cox, 1964). Ils proposent une famille de transformations indexées par un paramètre, noté $\lambda$ :

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & if \lambda \neq 0 \\ log(x) & if \lambda = 0 \end{cases} \tag{3.8}$$

En plus de la transformation logarithme, cette famille peut identifier la transformation carrée ($\lambda = 2$), la racine carrée ($\lambda = 0,5$), l'inverse ($\lambda = -1$) et d'autres entre ces valeurs. Le paramètre $\lambda$ est estimé en utilisant l'estimation du maximum de vraisemblance du prédicteur. Cette procédure serait appliquée indépendamment à chaque donnée de prédicteur contenant des valeurs supérieures à zéro.

# Chapitre 4

# Algorithmes d'apprentissage automatique

Ce chapitre a pour objectif de décrire les algorithmes d'apprentissage automatique qui sont repris dans le mémoire. Nous étudierons plusieurs modèles de classification allant du linéaire au non-linéaire. Enfin, nous conclurons ce chapitre par une présentation des techniques de re-échantillonage pour notre problématique (i.e. les données déséquilibrées) et des métriques d'évaluation de modèles.

L'ensemble des algorithmes et des techniques sera résumé dans ce chapitre avec les idées et les formules mathématiques principales. Pour une explication complète des mathématiques derrière les modèles utilisés dans cette étude, nous nous référons au lecteur à (Hastie, Tibshirani, & Friedman, 2009), et aux références utilisées dans le texte.

Nous souhaitons introduire les notations mathématiques utilisées dans ce chapitre. Supposons un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. On considère une variable vectorielle $X : \Omega \to \mathbb{R}^p$, avec un nombre entier $p$, et une variable de réponse $Y : \Omega \to \{0, 1\}$. Considérons un ensemble de données i.i.d (indépendantes et identiquement distribuées) $\mathcal{D}_n = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\}$. L'objectif (dans la plupart des algorithmes étudiés dans la suite) est d'estimer une fonction $f : \mathbb{R}^p \to [0, 1]$ telle que $f$ désigne la probabilité d'une observation labelisée 1.

## 4.1 Modèles de classification linéaire

Cette section présente les modèles linéaires qui sont très connus dans le monde statistique. Ces modèles sont largement utilisés dans la science actuarielle grâce à sa simplicité d'application et d'interprétation.

### 4.1.1 Régression logistique

La régression logistique est un modèle statistique qui utilise une fonction logistique pour estimer une variable dans l'intervalle $[0, 1]$. Il peut donc modéliser la probabilité conditionnelle d'une variable de réponse binaire. On considère un problème de classification binaire comme

suit :

$$\mathbb{P}(Y = 1|X) = f(X) = \sigma(\mathbf{w}^T X) \tag{4.1}$$

Où $\mathbf{w} \in \mathbb{R}^p$ le vecteur des paramètres et $f$ la fonction latente. Noter que sans perte de généralité, nous n'incluons pas d'intercept dans les modèles présentés dans ce chapitre. Cependant, il est considéré dans notre résultat numérique. De plus, $\sigma$ est la fonction sigmoïde :

$$\sigma(s) = \frac{1}{1 + e^{-s}}$$

Les paramètres du modèle logistique sont résolus en maximisant la fonction de vraisemblance. Ce dernier est équivalent à minimiser le logarithme négatif de vraisemblance. Pour un ensemble de données i.i.d (indépendantes et identiquement distribuées) $\mathcal{D}_n$, notons $\mathbb{P}(y_i = 1|\mathbf{x}_i) = p_i$, nous souhaitons minimiser la fonction de perte suivante :

$$\begin{aligned} L(\mathbf{w}) &= -\ln \mathbb{P}(Y|\mathbf{w}, X) = -\ln \prod_{i=1}^{n} \mathbb{P}(y_i|\mathbf{w}, \mathbf{x}_i) \\ &= -\ln \prod_{i=1}^{n} \mathbb{P}(y_i = 1|\mathbf{w}, \mathbf{x}_i)^{y_i} \left(1 - \mathbb{P}(y_i = 1|\mathbf{w}, \mathbf{x}_i)\right)^{1-y_i} \\ &= -\sum_{i=1}^{n} \left[y_i \ln p_i + (1 - y_i) \ln(1 - p_i)\right] \end{aligned} \tag{4.2}$$

L'estimateur de $\mathbf{w}$ est $\hat{\mathbf{w}}$ qui minimise $L(\mathbf{w})$. Contrairement au modèle de régression linaire qui a une solution analytique, il n'existe plus une telle solution pour le modèle logistique. La solution du modèle est donc trouvée par l'aide des algorithmes d'optimisation itératif tels que la méthode de descente de gradient stochastique ou l'algorithme Newton–Raphson.

### 4.1.2 Régression logistique pénalisée

Dans le cas de la régression logistique, Greene (2008) montre que la variance de l'estimateur de $\mathbf{w}$ est donnée par :

$$Var(\hat{\mathbf{w}}) = \left[\sum_{i=1}^{n} p_i(1 - p_i)\mathbf{x}_i^T \mathbf{x}_i\right]^{-1} \tag{4.3}$$

La grande valeur de la variance de l'estimateur n'est pas bonne pour le modèle car elle donne un impact défavorable sur la signification des variables explicatives ainsi que sur les tests d'hypothèse de validation du modèle (notons que le modèle logistique est paramétrique). Ainsi, pour éviter la grande variance de l'estimateur qui se produit lors de la grande multicolinéarité entre les variables indépendantes, on rajuste une pénalisation à la fonction de perte (4.2). Les 3 termes de pénalisation les plus connus comprennent Ridge, Lasso et Elastic Net.

La pénalité ridge est sous la forme de la somme des paramètres au carré, tandis que la pénalité Lasso est la somme des paramètres absolus. Enfin, L'eslastic net combine les deux pénalités précédentes. Plus précisement, les estimateurs des paramètres du modèle logistique en appliquant ces trois pénalités sont décrits respectivement par :

$$\hat{\mathbf{w}}_{Ridge} = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}} - \left[\sum_{i=1}^{n}(y_i \ln p_i + (1-y_i)\ln(1-p_i) + \lambda \sum_{j=1}^{p}(w_j)^2\right] \tag{4.4}$$

$$\hat{\mathbf{w}}_{Lasso} = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}} - \left[\sum_{i=1}^{n}(y_i \ln p_i + (1-y_i)\ln(1-p_i) + \lambda \sum_{j=1}^{p}|w_j|\right] \tag{4.5}$$

$$\hat{\mathbf{w}}_{EN} = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}} - \left[\sum_{i=1}^{n}(y_i \ln p_i + (1-y_i)\ln(1-p_i) + \lambda \sum_{j=1}^{p}|w_j| + \lambda' \sum_{j=1}^{p}|w_j|^2\right] \tag{4.6}$$

où $w_i$ dénote l'$i$-ième élément du vecteur $\mathbf{w}$, et $\lambda$ et $\lambda'$ sont les poids de pénalisation. Plus le poids de pénalisation est grand, plus les valeurs absolues des paramètres sont petites. D'ailleurs, la pénalité Lasso peut ramener les paramètres estimés à zéro. Cela n'arrive pas numériquement pour la pénalité Ridge. C'est pourquoi, en statistique, la régression Lasso est référencée pour la sélection des variables. De plus, $\lambda$ intervient ici lors de la minimisation, et non pas a posteriori. Il est donc un hyper-paramètre qui doit être optimisé (tuned en anglais) avant la phase d'optimisation du modèle.

Quelques commentaires plus généraux sur la façon dont les trois régressions se comparent :

- Souvent, ni l'un ni l'autre n'est globalement meilleur.
- Le Lasso peut mettre certains coefficients à zéro, effectuant ainsi une sélection de variables, contrairement à la régression Ridge.
- Le Lasso a tendance à bien fonctionner s'il y a un petit nombre de paramètres significatifs et que les autres sont proches de zéro (c'est-à-dire quand seuls quelques prédicteurs influencent réellement la variable de réponse).
- Le Ridge fonctionne bien lorsque la plupart des variables ont un impact sur la réponse.
- Enfin, l'Elastic Net est la combinaison de deux pénalités précédentes. Cependant, cela ne signifie pas que l'Elastic Net est supérieur aux deux autres, il est seulement une variante qui peut être plus adaptée (ou pas du tout) sur certains problèmes.

## 4.2 Modèles de classification non-linéaire

Dans cette section, nous allons présenter les modèles de classification non-linéaires. Pour ces modèles, la surface de réponse n'est plus un opérateur linéaire de variables explicatives.

En général, les variables explicatives seront projetées dans un autre espace par transformation non-linéaire, puis une formule linéaire pourrait être mise en œuvre pour les caractéristiques transformées (non pas les caractéristiques originales).

### 4.2.1 K plus proches voisins

La méthode de K plus proches voisins (KNN) utilise le voisinage géographique, i.e. les $K$ points les plus proches d'une observation pour prédire sa classe. La « proximité » est déterminée par une métrique de distance, et le choix de la métrique dépend des caractéristiques du prédicteur. Le choix par défaut est la distance euclidienne donnée par :

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{(x_i^1 - x_{i'}^1)^2 + ... + (x_i^p - x_{i'}^p)^2}$$

avec $x_i^k$ indique le $k$-ième élément d'observation $\mathbf{x}_i$.

Pour toute métrique de distance, il est important de se rappeler que les échelles de mesure d'origine des prédicteurs affectent les calculs de distance. C'est à dire qu'un prédicteur à grand échelle contribue un impact important à la distance, même si ce prédicteur ne influence pas actuellement la réponse. (Kuhn & Johnson, 2016). Pour surmonter cet enjeu, tous les variables explicatives (même la variable binaire) doivent être standardisées avant la mise en œuvre de KNN.

La classification de KNN pour un nouveau point se base sur l'ensemble des points déjà observés et ne nécessite aucun modèle (*model-free*). Ainsi, le KNN n'a pas de phase d'apprentissage et la phase de prédiction d'un nouveau point est implémentée par deux étapes :

- Calculer la distance $d$ entre $\mathbf{x}_i$ et chaque observation dans la base d'apprentissage, puis chercher une base des K observations les plus proches en distance de $\mathbf{x}_i$, notée $\mathcal{A}$.
- Estimer la probabilité conditionnelle de turnover de l'observation $\mathbf{x}_i$ par la fraction des points dans $\mathcal{A}$ qui ont le taux de turnover observé.

$$\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{K} \sum_{r \in \mathcal{A}} \mathbb{1}(y_r = 1) \tag{4.7}$$

où $\mathbb{1}$ est la fonction de l'indicateur.

L'estimation (4.7) tient compte que les $K$ points les plus proches ont la même contribution ou le même poids de vote sur le label du point considéré. Ce dernier peut devenir inflexible sur certaines problématiques. Ainsi, un variante du KNN considérant la pondération est introduite.

Son estimation de la probabilité conditionnelle est donnée comme suit :

$$\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = \frac{\sum_{r \in \mathcal{A}} w_r^i \mathbb{1}(y_r = 1)}{\sum_{r \in \mathcal{A}} w_r^i} \tag{4.8}$$

avec $w_r^i = 1/d(\mathbf{x}_i, \mathbf{x}_r)$.

Le KNN avec la pondération a été utilisé dans l'étude.

### 4.2.2 Machines à vecteurs de support

Les machines à vecteurs de support (SVM) sont une classe de modèles statistiques développés pour la première fois au milieu des années 1960 par Vladimir Vapnik. Au cours des dernières années, le modèle a considérablement évolué pour devenir l'un des outils disponibles d'apprentissage automatique les plus flexibles et les plus efficaces disponibles (Kuhn & Johnson, 2016). Une machine à vecteurs de support construit un « hyperplane » pour réaliser une séparation de classe. Intuitivement, le modèle est performant quand (i) les points segmentés dans le même groupe ont la même vraie réponse et (ii) l'hyperplane (la ligne dans le cas où on a deux variables indépendantes) considéré comme une barrière constitue des bonnes distances aux points les plus proches de cet hyperplane (la marge). Pour cette raison, le SVM est également appelé le classificateur de marge maximale (Kuhn & Johnson, 2016).

En général, le SVM est un modèle de noyau avancé avec les mathématiques complexes derrière, par exemple, la projection des données dans l'espace de Hilbert qui généralise la notion d'espace euclidien (dont la distance est utilisé dans le KNN), ou la propriété d'optimisation convexe importante. Dans la limitation de ce mémoire, nous ne présentons que la motivation et la forme mathématique du modèle.

Un exemple du modèle SVM avec 2 variables explicatives est affiché dans le graphique suivant :

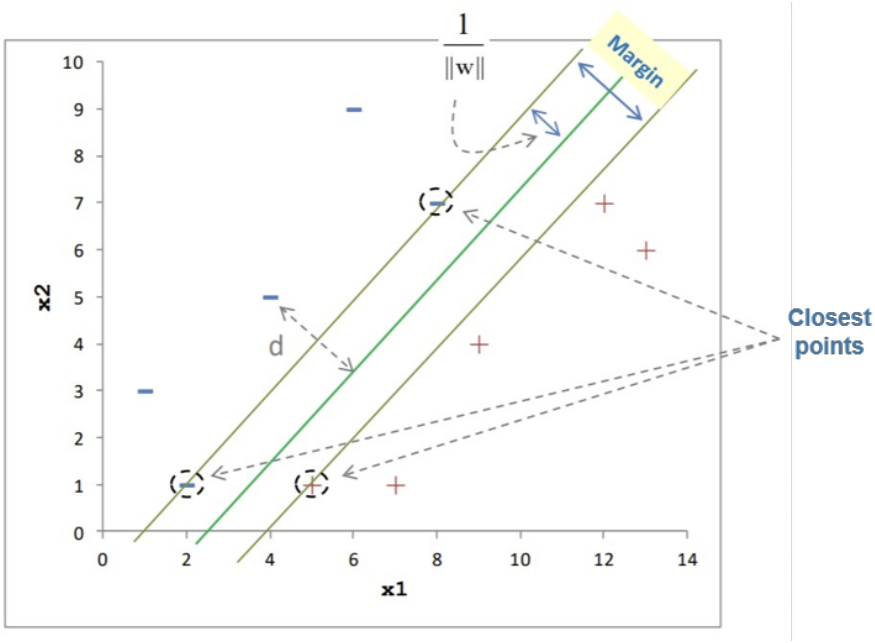



FIGURE 4.1 – (*Source : Cours de SVM de l'université de Lyon 2*) Exemple du modèle SVM linéaire avec deux variables explicatives

Sans perte de généralité, nous supposons que la réponse $Y$ maintenant est labelisée par $\{-1, 1\}$. Considérons un hyperplane $p$ comme suit :

$$p(X) = \mathbf{w}^T \phi(X) + b \tag{4.9}$$

Où $\mathbf{w}$ est un vecteur de paramètre, $b$ est le paramètre d'intercept et $\phi$ dénote une transformation d'espace de $X$. Dans le cas où $\phi$ est la fonction d'identité par point (i.e. $\phi(X) = X$), il s'agit d'un modèle linéaire.

La motivation du modèle est de trouver l'hyperplane $p$ qui satisfait que, pour chaque observation $i$, les signes de $p(\mathbf{x}_i)$ et du vrai label sont identiques. En résumé, comme illustré dans la figure 4.1, nous allons chercher un hyperplane (une ligne lors de deux variables explicatives) qui non seulement satisfait la condition de signe précédente mais aussi maximise la marge mesurée par les points les plus proches (de cet hyperplane). De plus, une point défini par (4.9), la distance de cet hyperplane à un point arbitraire $i$ est égale à :

$$d_i = \frac{\left|\mathbf{w}^T \phi(\mathbf{x}_i) + b\right|}{\|\mathbf{w}\|}$$

Où $\|\mathbf{x}\|$ est la norme euclidienne du vector $\mathbf{x}$. Alors, pour l'ensemble $\mathcal{D}_n$, la solution de ce modèle est la suivante :

$$\underset{\mathbf{w},b}{\operatorname{argmax}} \left\{ \frac{1}{\|\mathbf{x}\|} \min_i \left|\mathbf{w}^T \phi(\mathbf{x}_i) + b\right| \right\} \tag{4.10}$$

Avec la minimisation décrit les points les plus proches. La solution directe de ce problème d'optimisation est complexe, nous allons donc le convertir en un problème équivalent qui est plus facile à résoudre. Pour ce faire, on change l'échelle des paramètres $\mathbf{w}$ et $b$ de sorte que :

$$\left|\mathbf{w}^T\phi(\mathbf{x}_i) + b\right| \geq 1 \qquad \text{pour chacun de } i = 1, \ldots, n$$

Par conséquent, en rajoutant la condition de signe, le problème devient :

$$\begin{aligned} &\frac{1}{2}\operatorname*{argmin}_{\mathbf{w}} \|\mathbf{w}\|^2 \\ &\text{t.q. } y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1, \forall i = 1, \ldots, n \end{aligned} \tag{4.11}$$

Où la formularisation par le factor $1/2$ et la norme carrée a pour objet de convertir en un problème d'optimisation quadratique standard. En utilisant la méthode du multiplicateur lagrangien, le problème dual de (4.11) est :

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i}^{n}\sum_{j}^{n} \alpha_i\alpha_j y_i y_j \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle \\ &\text{t.q. } \begin{cases} \alpha_i \geq 0, \forall i \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \end{aligned} \tag{4.12}$$

Afin d'accélérer la computation de (4.12), on utilise l'astuce du noyau (*kernel trick* en anglais) en définissant :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$$

En choisissant le noyau linéaire, $K$ réduit à

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle\mathbf{x}_i, \mathbf{x}_j\rangle$$

La fonction de noyau la plus connue qui est aussi appliquée dans notre partie numérique est celle de base radiale :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \tag{4.13}$$

avec $\gamma$ est un hyper-paramètre à faire *tuning*. Dans notre numérique, $\gamma$ est mis, par défaut du package Sklearn de Python, à $1/p$ où $p$ est la dimension de $X$.

### 4.2.3 Réseau de neurones

Nous passons au dernier algorithme d'apprentissage automatique non-linéaire dans cette section. Cela est le réseau de neurones dont la conception est inspirée du fonctionnement des neurones biologiques comme illustré dans la figure 4.2. Dans le monde du *machine learning*, le

réseau de neurones est très connu et largement utilisé grâce à sa puissance. Ce dernier est testé sur une vaste expérience pratique et éprouvé théoriquement par le théorème d'approximation universelle. Plus précisément, ce théorème indique qu'un réseau d'une seule couche cachée contenant un nombre fini de neurones peut approximer n'importe quelle fonction continue. Dans la suite, nous présentons un réseau de neurones de deux couches cachées qui est utilisé dans notre étude (cf. Figure 4.2).

Pour $p, u$ entiers positifs et les fonctions scalaires par élément $\sigma$ et $\gamma$, soit $f$ est une fonction qui prend un vecteur $x \in \mathbb{R}^p$ en entrée et renvoie un vecteur $t \in [0,1]^2$ à travers la transformation séquentielle suivante :

$$\begin{aligned} a^0 &= z^0 = x \\ a^1 &= \sigma(z^1) = \sigma\left(w^1 a^0 + b^1\right) \\ a^2 &= \sigma(z^2) = \sigma\left(w^2 a^1 + b^2\right) \\ t &= \gamma(z^3) = \gamma\left(w^3 a^1 + b^3\right) \end{aligned} \tag{4.14}$$

Cette fonction est paramétrée par les poids $w^1 \in \mathbb{R}^{p\times u}, w^2 \in \mathbb{R}^{u\times u}$ et $w^3 \in \mathbb{R}^{u\times 1}$, et par les biais $b^1 \in \mathbb{R}^u, b^2 \in \mathbb{R}^u$ et $b^3 \in \mathbb{R}$. Alors, nous notons $\theta$ l'ensemble des paramètres du modèle.

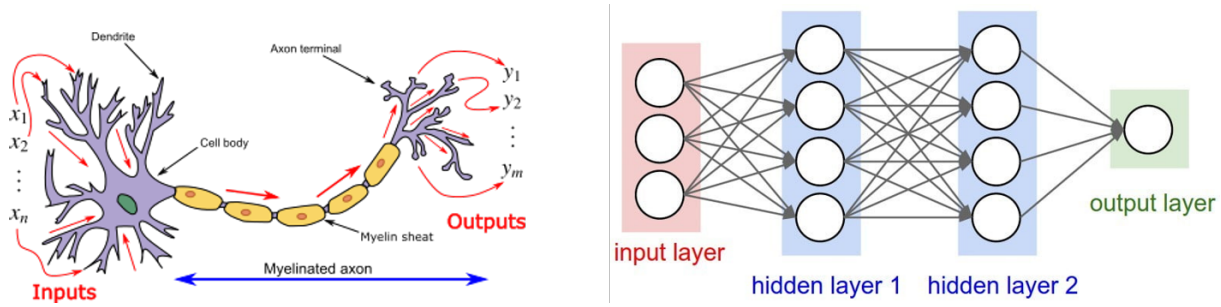



FIGURE 4.2 – (*Source : Wikipedia*) A gauche - Illustration des neurones biologiques, à droite - Réseau de neurones de deux couches cachées

Ceci est un réseau de classification de deux couches cachées contenant $u$ neurones et activé par $\sigma$. Dans (4.14), $z^i$ et $a^i$ sont appelé les sorties des couches cachées pré-activées et activées, respectivement. La fonction d'activation $\sigma$ peut être une fonction non-linéaire arbitraire, celle utilisée dans ce mémoire est le ReLU (*Rectified Linear Unit*) (définie par $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$). Tandis que la fonction $\gamma$ doit être la fonction sigmoïde pour que la sortie $t$ soit en fait la probabilité estimé de la classe 1.

Pour l'échantillon $\mathcal{D}_n$, similaire au modèle logistique, le réseau de classification entraîne une fonction de perte logarithmique suivante :

$$L_{NN}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln f_\theta(\mathbf{x}_i) + (1-y_i)\ln(1-f_\theta(\mathbf{x}_i))\right] \tag{4.15}$$

Alors, (4.15) est minimisée par la descente de gradient. Ceci est un algorithme d'optimisation local dans lequel le paramètre $\theta$ est ajusté de manière itérative dans le sens du gradient négatif de la fonction objectif. Pour un taux d'apprentissage $lr$ positif, la mise à jour à l'étape $k$ peut être écrite comme

$$\theta_{k+1} = \theta_k - lr\nabla_\theta L_{NN}(\theta_k)$$

Même si les calculs de gradient peuvent être simplement mis en œuvre par différenciation automatique dans la pratique (Paszke et al. 2017), le calcul du gradient à partir d'un grand ensemble de données reste coûteux en temps de calcul. De plus, la descente de gradient peut facilement se coincer à un point minimum local indésirable. Pour ces raisons, la descente de gradient est pratiquement remplacée par la descente de gradient stochastique (SGD). Le SGD (vanille) peut être vu comme une approximation stochastique de la descente de gradient, car au lieu d'utiliser l'ensemble de données, il calcule le gradient à partir d'un sous-ensemble aléatoire des données (ou même un seul échantillon). Contre la descente de gradient, le SGD réalise un calcul plus rapide à chaque itération dans le commerce pour un taux de convergence plus faible. Du point de vue mathématique, cette technique n'est rien d'autre que le théorème de dérivation des fonctions composées (parfois appelé règle de dérivation en chaîne ou règle de la chaîne).

Il faut mentionner que l'entraînement du réseau de neurones est effectué à l'aide de la technique de rétro-propagation qui nous permet de calculer efficacement les gradients par rapport aux paramètres dans les couches cachées. Les lecteurs qui souhaitent voir en détail l'optimisation et la rétro-propagation du réseau de neurones peuvent se référencer le livre (Goodfellow et al. 2016).

Bien que ce modèle soit puissant, nous n'obtenons pas un résultat numérique favorable par le réseau de classification (cf. le chapitre 5). Nous pensons donc que le réseau de neurones (ou au moins le réseau standard) n'est pas adapté au problème de classification déséquilibrée où la présence d'une classe est très déséquilibrée (i.e. environ 4 %).

## 4.3 CART et Modèles d'ensemble

### 4.3.1 CART

L'arbre de décision est une approche couramment utilisée en *machine learning* pour la modélisation de classification grâce à sa simplicité et de sa capacité d'interprétation. Un arbre de décision est une représentation visuelle d'un algorithme de classification de données suivant

différents critères qu'on appellera décisions (ou noeuds). L'ensemble des notations dans un modèle est donné par le graphique suivant :

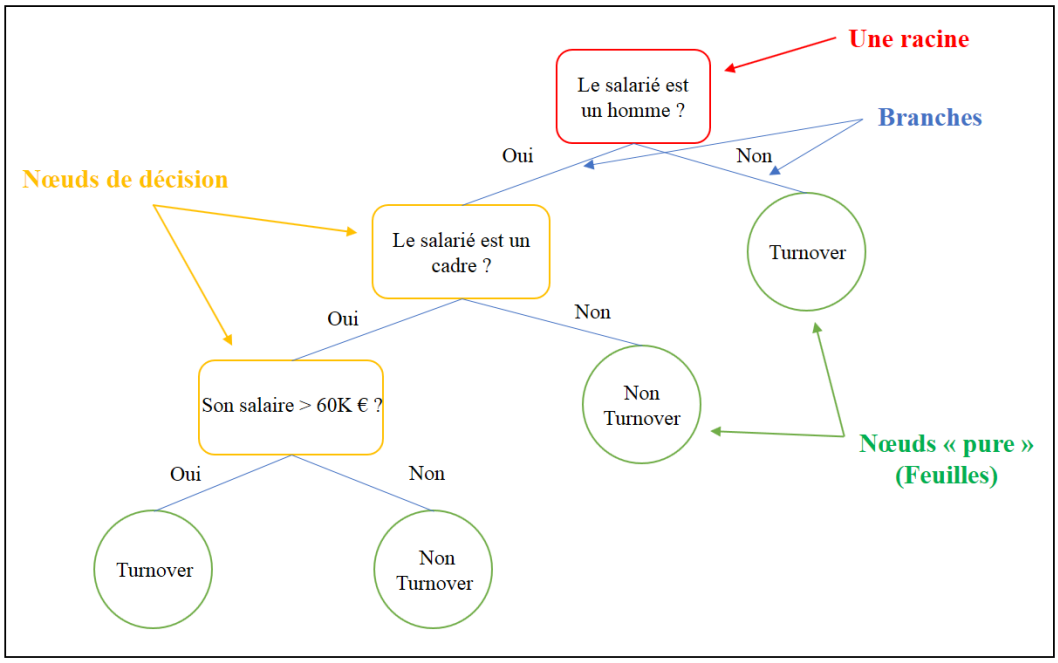



FIGURE 4.3 – Exemple d'un arbre de décision

- Une racine : contient l'ensemble de la population à segmenter. C'est le point de départ ;
- Des branches : contiennent les règles de division qui permettent de segmenter la population ;
- Des nœuds de décision : contiennent les sous-populations homogènes (sur leurs caractéristiques et la réponse) créées, fournissent l'estimation de la quantité d'intérêt ;
- Des nœuds « pure » ou des feuilles : indiquent la classe résultante.

L'algorithme CART doit décider automatiquement des variables et des nœuds de segmentation, ainsi que de la topologie (forme) que l'arbre doit avoir.

Pour l'ensemble $\mathcal{D}_n$, la probabilité conditionnelle d'intérêt est :

$$p(\mathbf{x}_i) = \mathbb{P}(y_i = 1|\mathbf{x}_i) \tag{4.16}$$

Supposons qu'il y une partition en M régions $R_1, R_2, ..., R_M$. Au nœud $m$ correspondant à la région $R_m$ avec $N_m$ observations, la proportion des observations de turnover est définie comme la suivante :

$$p_m = \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} \mathbb{1}_{y_i=1} \tag{4.17}$$

Dans le modèle de CART pour la classification binaire, l'indice Gini est couramment utilisé afin de mesure l'impureté du nœud $m$ et est donné par :

$$G_m = 2p_m(1 - p_m) \tag{4.18}$$

Supposons que $R_1$ et $R_2$ sont les régions des deux feuilles, l'algorithme calcule la partition optimale pour laquelle la valeur de $G_{R_1} + G_{R_2}$ est **minimum**, i.e. à chaque étape $m$, la division de « une région supérieure » ($R$) en « deux régions inférieures » ($R_1$ et $R_2$ correspondant à la construction récursive de l'arbre) **maximise** la différence de mesure d'impureté du nœud (appelée également la fonction d'hétérogénéité ou la déviance) :

$$\hat{\Delta}_{\{\mathbf{x}^j, 1<=j<=p\}} G_{R \to R_1 + R_2} = G_R - \left( \frac{N_{R_1}}{N_R} G_{R_1} + \frac{N_{R_2}}{N_R} G_{R_2} \right) \tag{4.19}$$

**L'algorithme de création d'un arbre sera résumé dans le graphique suivant :**

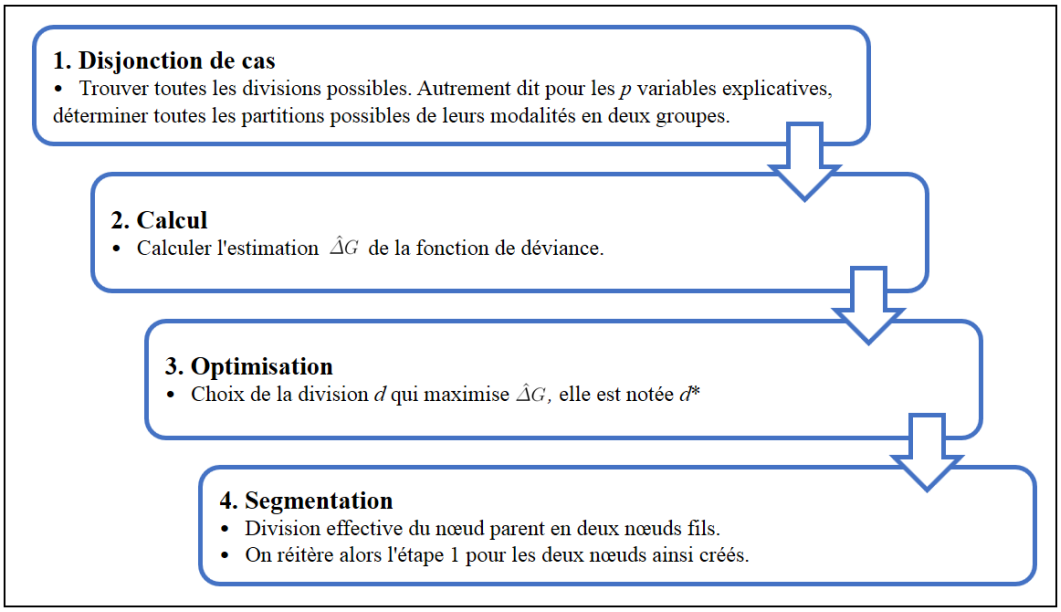



FIGURE 4.4 – Algorithme de création d'un arbre

L'algorithme s'arrête le processus de division lors qu'un nœud est terminal si :

- la région est homogène (une seule modalité) ;
- il n'y a pas de partitions autorisées concernant la règle algorithmique de diminution du critère de variance $\Delta G$
- le nombre d'observations dans la région $N_{Cut}$ (ou dans les sous régions $N_{Size}$) est inférieur à un seuil donné alors Pas de partition autorisée (paramètres d'algorithme).

- Les paramètres de tuning : $(N_{Cut}, N_{Size})$

**Le modèle de CART apporte plusieurs avantages comme les suivants :**

- Un avantage important du modèle de classification CART est l'interprétabilité de l'arbre, qui décrit totalement la partition qui a été faite de l'espace dans lequel la variable à expliquer prend ses valeurs. Néanmoins, il est à noter que lorsque la profondeur de l'arbre est trop grande, la lisibilité de l'arbre devient plus difficile.
- Par rapport à d'autres algorithmes, le modèle CART nécessite moins d'efforts pour la préparation des données. Par ailleurs, les valeurs manquantes dans les données n'affectent pas le processus de construction d'un arbre de décision dans une mesure considérable.
- Enfin, la classification CART n'est pas paramétrique et nécessite donc de ne faire aucune hypothèse : sur la loi de la variable à expliquer, sur l'indépendance des variables explicatives, et sur les modalités à effectuer au sein des variables explicatives catégorielles.

**Cependant, il existe quelques inconvénients qui sont détaillés comme les suivants :**

- Le premier est leur manque de robustesse. Un petit changement dans les données peut résulter un changement important dans la structure de l'arbre de décision entraînant une instabilité.
- Le risque de sur-apprentissage est considérablement élevé avec les arbres de décision et ils ont tendance à rester coincés dans les minimas locaux. Cela peut détruire l'expérience d'apprentissage automatique.
- Enfin, le modèle étant non probabiliste, il ne peux pas proposer directement des intervalles de confiance associés aux prédictions sans avoir l'aide des méthodes de bootstrap.

### 4.3.2 Forêts aléatoires

Comme montré dans la partie précédente, l'arbre de décision généré par l'algorithme CART offre une grande flexibilité dans la modélisation de classification, notamment en captant des effets non linéaires et en introduisant des interactions. Cependant, avec cette méthode, le risque de sur-apprentissage augmente également. En effet, les arbres CART sont généralement instables et peuvent varier considérablement par rapport à juste un changement mineur dans les données.

Afin de pallier ce défaut, des techniques d'agrégation de plusieurs arbres sont généralement utilisées pour générer un modèle plus robuste. Ces méthodes utilisent la technique de « Bootstrap » (i.e. le tirage avec remise d'un échantillon) afin de créer une multitude d'arbres.

Supposons qu'il y a une suite $(\Phi_k)_{k=1,\dots,B}$ de B arbres obtenus par l'agrégation et l'arbre final est défini par $\Phi_k = \sum_{k=1}^{B} \Phi_k$. L'agrégation permet de diminuer la variance de l'estimation. Par conséquent, si les arbres $\Phi_k$ sont deux à deux corrélés par $\rho$ et ils ont une variance $\sigma^2$, la variance de l'arbre final est donnée par :

$$\begin{aligned} Var[\Phi] &= Var\left[\frac{1}{B}\sum_{k=1}^{B}\Phi\right] = \frac{1}{B^2}Var\left[\sum_{k=1}^{B}\Phi\right] = \frac{1}{B^2}\sum_{k=1}^{B}\left(\sigma^2 + \sum_{k'=1,k'\neq k}^{B}\rho\sigma^2\right) \\ &= \frac{\sigma^2}{B^2}\sum_{k=1}^{B}\Big(1+(B-1)\rho\Big) = \frac{\sigma^2}{B}\Big(1+(B-1)\rho\Big) = \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{B} \end{aligned} \tag{4.20}$$

La variance décroît en fonction de nombre des arbres « plantés » : $Var[\Phi] \xrightarrow[B\to\infty]{} \rho\sigma^2$.

Parmi les méthodes d'agrégation, l'algorithme de forêts aléatoires (RF) est utilisé les plus souvent car il permet d'introduire plus d'indépendance entre les arbres construits, et donc de réduire la variance ci-dessus.

**L'algorithme du modèle de forêts aléatoires est affiché dans le graphique suivant :**

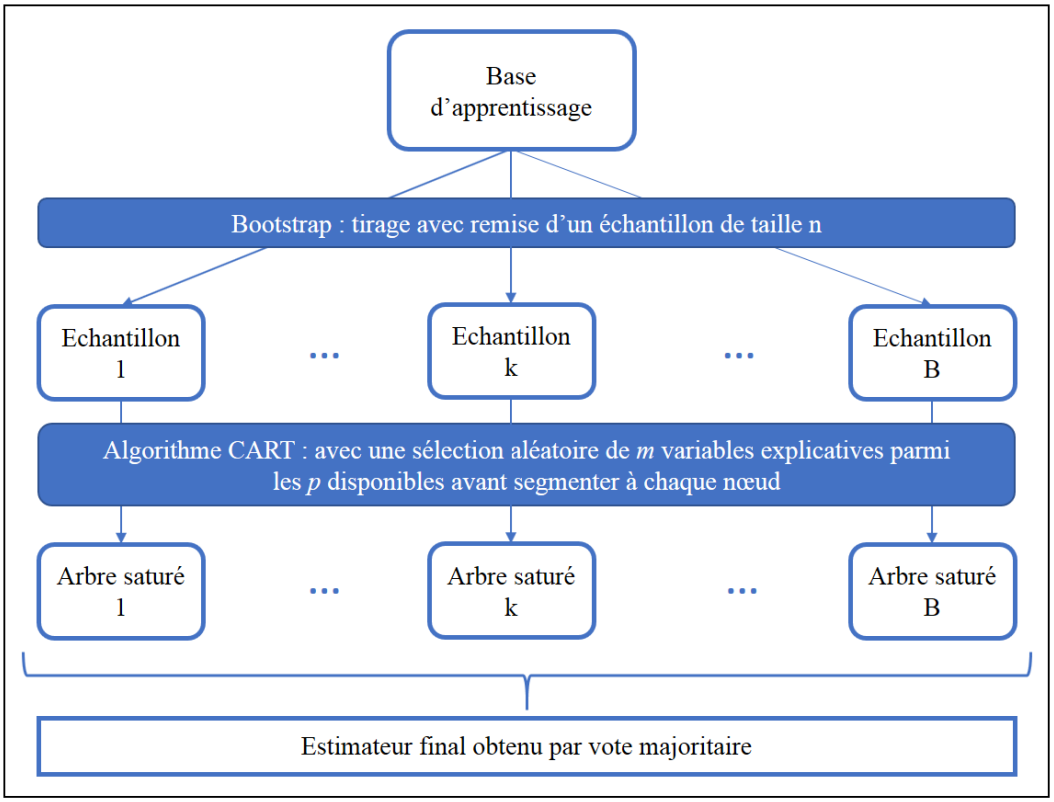



FIGURE 4.5 – Algorithme de Forêts aléatoires

Dans chaque échantillon créé par Boostrap, certain nombre $m$ de variables explicatives est

aléatoirement tiré et l'arbre sera construit au travers de ce sous-échantillon. L'estimateur final est calculé par la moyenne des différents estimateurs pour la régression et par le vote à la majorité pour la classification. Cette technique permet de rendre les arbres « plantés » plus « indépendants », la corrélation $\rho$ entre les variables est diminuée, d'où la décroissance de la variance du l'estimateur final.

Par défaut, le nombre de variables explicatives tirées aléatoirement est $m \sim \sqrt{p}$ pour un arbre de classification.

Soit $\hat{y}_i$ l'estimateur obtenu pour l'individu i par un CART maximal. On construit B arbres CART par l'approche de forêts aléatoires. Pour chaque observation $i$, l'estimateur forêts aléatoires vaut :

$$\hat{y}_i^{RF} = \arg\max_{k=0,1}(\#\hat{y}_i^{CART} = k) \tag{4.21}$$

### 4.3.3 Extreme Gradient Boosting

(Friedman, J., Hastie, T., & Tibshirani, 2000) ont travaillé pour fournir un aperçu statistique de l'algorithme AdaBoost. Pour le problème de classifica tion, ils ont montré qu'il pouvait être interprété comme un modèle additif progressif par étape qui minimise une fonction de perte exponentielle (Kuhn & Johnson, 2016). Ce cadre a conduit à des généralisations algorithmiques telles que Real AdaBoost, Gentle AdaBoost et LogitBoost. Ces généralisations ont été placées dans un cadre unificateur appelé machine de Gradient Bossting. Les principes de base de Gradient Bossting sont les suivants : étant donné une fonction de perte (par exemple, shrinkage) et un apprenant faible (par exemple, des arbres de classification), l'algorithme cherche à trouver un modèle additif qui minimise la fonction de perte. Le gradient (par exemple, résiduel) est calculé, et un modèle est ajouté au modèle précédent, et la procédure se poursuit pendant un nombre d'itérations spécifié par l'utilisateur (Kuhn & Johnson, 2016).

Un classificateur faible est un classificateur dont le taux d'erreur n'est que légèrement meilleur par rapport à une estimation aléatoire. Le but de Boosting est d'appliquer séquentiellement l'algorithme de classification faible à des versions modifiées des données, produisant ainsi une séquence de classificateurs faibles $C_m(x), m = 1, 2, ..., M$. Supposons que l'espace de toutes les observations de variables explicatives est divisé en $J$ régions disjointes $R_{j,j=1,2,...,J}$ représentées par les nœuds terminaux de l'arbre.

La fonction de classification globale ou l'arbre est donnée par :

$$T(X, \Theta) = \sum_{j=1}^{J} \gamma_j \mathbb{1}_{X \in R_m} \tag{4.22}$$

avec $\gamma_j \in \{0, 1\}$ une modalité constante pour chaque région $R_m$ et $\Theta = \{R_j, \gamma_j\}_1^J$. Les paramètres sont trouvés en minimisant le risque empirique :

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in R_j} L(y_i, \gamma_j) \tag{4.23}$$

Étant difficile de trouver les valeurs optimales du problème d'optimisation ci-dessus, il est parfois nécessaire d'approcher (7.25) par un critère plus lisse et plus pratique pour optimiser le $R_j$ :

$$\tilde{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \tilde{L}(y_i, T(\mathbf{x}_i, \Theta)) \tag{4.24}$$

Pour résoudre le problème d'optimisation, une stratégie typique est d'utiliser l'indice Gini décrite dans la sous section 4.3.1 en construction de l'arbre (identifiant / optimisant la $R_j$).

Le modèle d'arbre boosté est une somme de ces arbres qui induisent de manière progressive :

$$f_M(X) = \sum_{m=1}^{M} T(X, \Theta_m) \tag{4.25}$$

À chaque étape de la procédure progressive, il faut résoudre :

$$\tilde{\Theta}_m = \underset{\Theta_m}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f_{m-1}(\mathbf{x}_i) + T(\mathbf{x}_i, \Theta)) \tag{4.26}$$

La technique de descente de gradient sera utilisée pour trouver les valeurs optimales, et c'est la raison pour la quelle on parle de « Gradient Boosting ». Outre la taille des arbres constituants, $J$, l'autre méta-paramètre du boosting de gradient est le nombre d'itérations boosting $M$. Chaque itération réduit généralement le risque d'entraînement $L(f_M)$, de sorte que pour $M$ assez grand ce risque peut être rendu arbitrairement petit. Cependant, un ajustement trop précis des données d'apprentissage peut entraîner un sur-apprentissage, ce qui dégrade le risque sur les prévisions futures.

Dans ce mémoire, l'algorithme de « Extreme Gradient Boosting » (XGB) a été retenu. Le XGB est généralement la méthode la plus performante parmi l'ensemble des algorithmes Boosting car il utilise la stratégie de régulation (techniques de « shrinkage ») afin de réduire le problème de sur-apprentissage. La mise en œuvre la plus simple de shrinkage dans le contexte du boosting est de mettre à l'échelle la contribution de chaque arbre d'un facteur $0 < \nu < 1$ lorsqu'il est ajouté à l'approximation courante.

$$f_m(X) = f_{m-1}(X) + \nu \sum_{j=1}^{J} \gamma_{jm} \mathbb{1}_{X \in R_m} \tag{4.27}$$

**L'algorithme de XGB :**

1. Initialiser $f_0(\mathbf{x}_i) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

2. Pour $M = 1, ..., M$ :

   (a) Pour $i = 1, ..., n$, calculer :

   $$r_{im} = -\Big[\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\Big]_{f=f_{m-1}}$$

   (b) Ajuster un arbre à $r_{im}$ donnant les régions terminales $R_{jm}, j = 1, ..., J_m$.

   (c) Pour $j = 1, ..., J_m$, calculer :

   $$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{\mathbf{x}_i \in R_{jm}} L(y_i, f_{m-1}(\mathbf{x}_i) + \gamma)$$

   (d) Mettre à jour

   $$f_m(\mathbf{x}_i) = f_{m-1}(\mathbf{x}_i) + \nu \sum_{j=1}^{J} \gamma_{jm} \mathbb{1}_{\mathbf{x}_i \in R_m})$$

3. Sortir $\hat{f}(\mathbf{x}_i) = f_M(\mathbf{x}_i)$.

Dans le problème d'optimisation (4.26), chaque observation constitue le même poids dans la fonction de perte totale. Cependant, on peut utiliser une fonction de perte pondérée pour être plus adapté au problème de classification déséquilibrée. En particulier, le poids des observations qui appartiennent à la classe minoritaire sera plus grand que celui des autres. Ce modèle est noté *Balanced XGB* dans notre résultat numérique.

### 4.3.4 Généralisation empilée - *Stacking model*

On conclut l'introduction des algorithmes de *machine learning* par une technique qui s'appelle généralisation empilée ou *Stacking* (Wolpert, David H., 1992). Le *Stacking* est une méta-modélisation qui nous permet d'agréger les estimations des différents algorithmes. En effet, la généralisation empilée fait partie d'apprentissage ensembliste comme la forêt aléatoire et le *boosting*. A noter que contrairement à ces derniers qui combinent les mêmes modèles naïfs (i.e. le CART), le *Stacking* peut composer les différents modèles (modèles de base). De plus, sa méthode d'agrégation ne se limite qu'à utiliser la moyenne des estimations des modèles de

base, mais peut être effectuée en entraînant un autre modèle (méta-modèle). Pour résumer, les étapes du *Stacking* s'articulent comme suit (cf. Figure 4.6) :

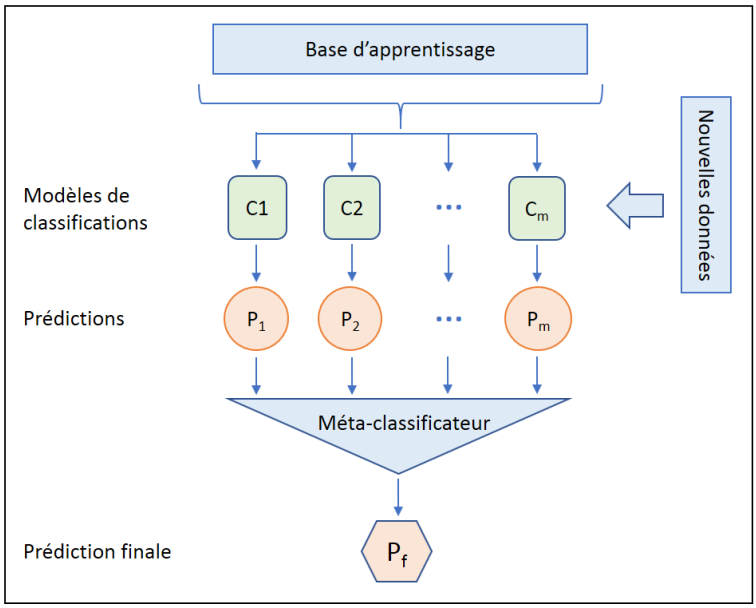



FIGURE 4.6 – Illustration graphique de la généralisation empilée

1. Apprendre divers modèles de base à partir des données apprentissage et puis produire leurs prédictions en utilisant les différents modèles de base.
2. Constituer un nouvel échantillon d'apprentissage depuis les prédictions dans l'étape 1. Plus précisément, la prédiction de chaque modèle de base est considérée comme une variable explicative dans le nouvel échantillon et la variable de réponse ne change pas par rapport aux données dans l'étape 1.

Enfin, voici quelques remarques sur la structure de la généralisation empilée :

- Modèles de niveau 0 (modèles de base) : les modèles s'adaptent aux données d'apprentissage avec les prédictions qui y sont compilées. Les modèles de base sont souvent complexes et diversifiés.
- Modèle de niveau 1 (méta-modèle) : le modèle qui apprend à combiner au mieux les prédictions des modèles de base. Le méta-modèle est souvent simple, fournissant une interprétation fluide des prédictions faites par les modèles de base. Pour la régression, le méta-modèle est souvent la régression linéaire. Pour la classification, le choix fréquent pour le méta-modèle est la régression logistique qui est également utilisée dans l'étude.

Le *Stacking* est conçu pour améliorer la performance de la modélisation, mais il ne garantit pas l'amélioration dans tous les cas. Cela dépend de la complexité du problème, du choix des

modèles de base et du fait que ces modèles de base soient suffisamment habiles et non corrélés dans leurs prédictions.

## 4.4 Techniques de re-échantillonage pour les données déséquilibrées

Il convient de répéter que le problème de classification déséquilibrée est un défi. Ainsi, à part la présentation des modèles de machine learning innovant, nous introduisons maintenant les techniques d'échantillonage particulières pour résoudre ce problème.

L'idée principale est de générer ou échantillonner un ensemble de données d'apprentissage dans lequel la réponse est labelisée de manière équilibrée. Pour ce faire, on va soit dupliquer aléatoirement les exemples de la classe minoritaire (cf. le sur-échantillonnage), soit sélectionner aléatoirement les exemples de la classe majoritaire qui seront supprimés de l'ensemble des données d'apprentissage (cf. le sous-échantillonnage). Les deux techniques sont illustrées dans la figure 4.7.

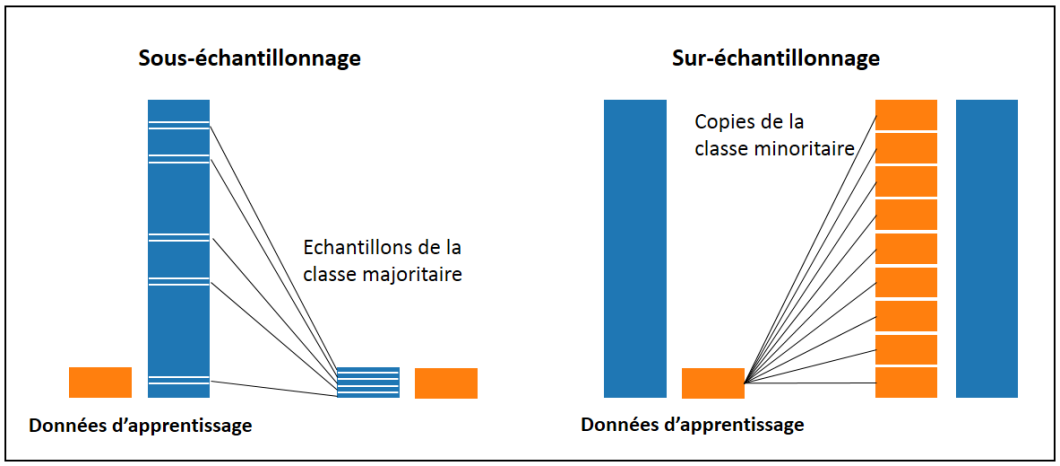



FIGURE 4.7 – Sous-échantillonage versus Sur-échantillonage

Apparemment, le sous-échantillonnage supprime beaucoup d'observations qui peuvent contenir les informations importantes, cela constitue donc un impact indésirable aux classificateurs. Alors qu'en reproduisant les observations de la classe minoritaire, le sur-échantillonage peut mieux recueillir les caractéristiques de cette classe, et provoquer probablement un problème de sur-apprentissage. Cependant, il n'y a pas encore (ou on n'a pas encore trouvé) les théories qui prouvent la justesse des deux techniques. Ainsi, afin de trouver le meilleur algorithme, nous allons tester les deux techniques en pratique.

Dans notre partie numérique, les classificateurs *Balanced Bagging Classifier* et *Balanced Random Forest Classifier* impliquent le sous-échantillonage est effectué à chaque étape d'échantillonage (*boostraping*). C'est à noter que les modèles sont différents par rapport aux modèles de *Bagging* et *Random Forest* avec les données sous-échantillonnées. Plus précisement, les premières techniques refont plusieurs fois le sous-échantillonage à partir d'ensemble de données originales, tandis que les deuxièmes techniques font un sous-échantillonage, puis sur le nouvel ensemble de données, ils refont le *boostraping* standard.

## 4.5 Métriques d'évaluation des modèles

Dans le contexte de la classification, les métriques telles que $RMSE$ et $R^2$ ne sont pas appropriées . Certaines mesures utiles pour la classification sont l'exactitude, la sensibilité, la spécificité, la valeur prédictive positive, l'exactitude équilibrée, le score F1, et la courbe *Receiver Operating Characteristic* (ROC) et *Area under the ROC Curve* (AUC).

Dans la pratique, l'exactitude est souvent choisie comme la métrique d'évaluation principale, cela est important lors du conflit des mesures pour comparer les modèles. Pourtant, ce choix n'est plus adapté dans le problème de classification déséquilibrées. Car, par exemple dans notre cas d'étude où la classe minoritaire ne se présente que 4% de l'ensemble de données, cela implique une prédiction toute simple qui rend toujours la classe majoritaire peut aussi avoir une grande exactitude. Pour cette raison, dans ce problème, la métrique principale doit être la courbe ROC et l'aire sous la courbe. Dans la suite, nous introduisons l'ensemble de toutes les métriques d'évaluation.

| | | Condition réelle | | | |
|---|---|---|---|---|---|
| | | Condition positive | Condition négative | | |
| Condition estimée | Condition positive estimée | Vrais positifs (VP) | Faux positifs (FP) | **PPV (Précision)** $\frac{VP}{VP+FP}$ | |
| | Condition négative estimée | Faux négatifs (FN) | Vrais négatifs (VN) | | |
| | | **Sensibilité** $\frac{VP}{VP+FN}$ | **Spécificité** $\frac{VN}{VN+FP}$ | **Exactitude** $\frac{VP+VN}{VP+VN+FP+FN}$ | **Exactitude équilibrée** $\frac{Sensitivité+Spécificité}{2}$ |
| | | | | **F1** $\frac{2*PPV*Sensitivité}{PPV+Sensitivité}$ | |

Tableau 4.1 – Métrique de confusion, y compris les mesures d'évaluation utilisées dans l'étude

Pour plus de clarification, le tableau 4.1 présente la matrice de confusion et ses mesures de base (i.e. VP, FP, FN et VN) ainsi que la relation avec les mesures d'évaluation. Ceux-ci seront expliqués ensuite.

### 4.5.1 Exactitude

L'exactitude (« Accuracy » en anglais) est l'une des mesures les plus simples. Il reflète avec l'interprétation la plus simple la concordance entre les classes observées et estimées. Cependant, dans les situations où la structure de la population et les coûts d'intérêt sont différents, l'exactitude peut ne pas mesurer les caractéristiques importantes du modèle. En outre, les fréquences naturelles de chaque classe doivent être prises en considération. Par exemple, dans le cas de l'étude, une règle simple stipulant que tous les salariés resteront dans l'entreprise se traduira déjà par une exactitude de 95,87 %. Bien que 95,87 % puissent être considérés comme élevés dans d'autres applications de modélisation prédictive, le taux d'exactitude semble inapproprié dans notre étude car nous sommes dans un cas de données déséquilibrées. La formule de l'exactitude est comme suivante :

$$Exactitude = \frac{\#\ observations\ correctement\ estimées}{Population\ totale} = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.28}$$

### 4.5.2 Sensibilité et Spécificité

Pour la classification des variables de réponse binaire, la sensibilité et la spécificité sont deux statistiques complémentaires qui peuvent être pertinentes. La sensibilité (« Sensitivity » en anglais) du modèle est définie comme le rapport entre le nombre d'observations de l'événement d'intérêt correctement estimées et le nombre d'observations ayant l'événement. La formule de la sensibilité est donnée comme suivante :

$$Sensibilité = \frac{\#\ obs.\ de\ l'événement\ d'intérêt\ correctement\ estimées}{\#\ observations\ ayant\ l'événement} = \frac{VP}{VP + FN} \tag{4.29}$$

La sensibilité est parfois considérée comme le taux vrai positif puisqu'elle mesure l'exactitude dans la population de l'événement d'intérêt. Inversement, la spécificité (« Specificity » en anglais) est définie comme le rapport entre les observations de non-occurrence de l'événement correctement estimées et toutes les observations sans l'événement. Le calcul de la spécificité est donnée comme suit :

$$Spécificité = \frac{\#\ obs.\ de\ non\text{-}occurrence\ de\ l'événement\ correctement\ estimées}{\#\ observations\ sans\ l'événement} = \frac{VN}{VN + FP} \tag{4.30}$$

### 4.5.3 Valeur prédictive positive

Tout comme la sensibilité et la spécificité, la valeur prédictive positive (« positive predictive value » en anglais) (PPV), également appelée précision, peut être calculée à l'aide de la matrice de confusion. Un aspect souvent négligé de la sensibilité et de la spécificité est qu'il s'agit de mesures conditionnelles. La sensibilité est le taux d'exactitude pour la population de l'événement uniquement (et la spécificité pour les non-événements). Intuitivement, si l'événement est rare, cela devrait être reflété dans la réponse. La PPV en tient compte et constitue une mesure inconditionnelle de la condition positive. La formule est suivante :

$$PPV = \frac{VP}{VP + FP} \tag{4.31}$$

### 4.5.4 Exactitude équilibrée

Étant donné que l'exactitude peut être une mesure de performance trompeuse (comme décrit dans la section sur l'exactitude), elle peut suggérer à tort une décision ou une généralisation. Pour éviter de rapporter une estimation d'exactitude optimiste, l'exactitude équilibrée (« Balanced Accuracy » en anglais) est introduite. L'exactitude équilibrée peut être définie comme l'exactitude moyenne obtenue sur l'une ou l'autre des classes. Sur la base d'une matrice de confusion, l'exactitude équilibrée est donnée par :

$$\textit{Exactitude équilibrée} = \frac{1}{2}(\textit{Sensibilité} + \textit{Spécificité}) \tag{4.32}$$

### 4.5.5 Le score F1

Le score F1 est une mesure de la précision d'un test pour la classification binaire. Il prend en compte à la fois la sensibilité et la PPV. Le score F1 peut être interprété comme une moyenne pondérée de ces deux mesures. Le score F1 tient compte des valeurs entre 1 et 0 où 1 est considéré comme le meilleur. La formule du calcul du score F1 est donnée comme ci-dessous :

$$F1 = 2\frac{\textit{Sensibilité} * PPV}{\textit{Sensibilité} + PPV} \tag{4.33}$$

Cependant, il est à noter que le score F1 ne prend pas directement en compte les vrais négatifs dans son calcul.

### 4.5.6 Courbe ROC-AUC

La courbe ROC est créée en évaluant les probabilités de classe pour le modèle sur un continuum de seuils. Pour chaque seuil candidat, le taux de vrais positifs qui en résulte (c'est-à-dire la sensibilité) et le taux de faux positifs (un moins la spécificité) sont tracés l'un par rapport à l'autre. Le seuil par défaut est de 50 %, qui est également le seuil utilisé pour indiquer les résultats. Ce seuil peut être modifié pour choisir un nouveau compromis de sensibilité et de spécificité, comme indiqué dans la section précédente.

La courbe ROC peut également être utilisée pour une évaluation quantitative du modèle. Un modèle parfait qui sépare complètement les deux classes aurait une sensibilité et une spécificité de 100 %. Dans ce cas, la courbe ROC serait graphiquement un angle droit au point (0, 1). L'aire sous la courbe ROC (« AUC » en anglais) pour un tel modèle serait 1. Le ROC d'un modèle complètement est une ligne de 45 dégrée et son AUC est donc égal à 0,5.

Un avantage de l'utilisation des courbes ROC pour caractériser les modèles est que, étant fonction de la sensibilité et de la spécificité, la courbe est insensible aux disparités dans les proportions de classes (Fawcett, 2006). Étant donné que des valeurs de seuil modifiées peuvent complètement changer les autres mesures décrites ci-dessus, les mesures de performance qui sont indépendantes des seuils de probabilité (comme l'AUC) sont susceptibles de produire des contrastes plus significatifs entre les modèles. C'est la raison pour laquelle cette métrique est choisie pour sélectionner le meilleur modèle dans l'étude.

# Chapitre 5

# Résultats numériques et Application sur les calculs actuariels

Dans ce chapitre, nous allons comparer les résultats des algorithmes d'apprentissage automatique ainsi que des techniques d'échantillonage introduits dans le chapitre précédent. Toutes les tentatives ont pour objet de trouver le meilleur estimateur du taux de turnover. En basant sur ce dernier, on va construire des tables de turnover attendues en projetant les hypothèses futures. Ensuite, l'application des nouvelles tables de turnover sur les calculs des engagements et les impacts associés seront effectués et analysés. Pour finir, nous résumerons des limites de notre étude.

## 5.1 Comparaison des algorithmes d'apprentissage automatique et Construction des tables de turnover

Cette section résume les résultats numériques obtenus par les algorithmes d'apprentissage automatique et présente ensuite la construction des nouvelles tables de turnover. Les études ont été réalisées sous l'aide du langage de programmation Python. Dans la limite de ce mémoire, nous ne allons pas présenter les détails des résultats de chaque modèle d'apprentissage automatique tels que : les paramètres et hyper-paramètres optimisés, les tests statistiques associés, et les graphiques ou les tableaux des métriques d'évaluation (ex. la métrique de confusion). A des fins de prédiction, seules les métriques d'évaluation principales ont été présentées pour comparer et trouver la meilleure estimation du taux de turnover.

### 5.1.1 Comparaison des algorithmes d'apprentissage automatique

Dans le tableau 5.1 suivant, nous rapportons l'ensemble des résultats de tous les modèles. Ces modèles sont entraînés sur l'échantillon d'apprentissage traité (cf. Sous-section 3.2.3), mais aucune technique d'échantillonage (cf. Section 4.4) est prise en application pour ce moment, puis les résultats dans le tableau 5.1 sont calculés sur l'échantillon de validation.

L'ensemble des métriques d'évaluation obtenues par les modèles d'apprentissage automatique est résumé comme suit :

| Modèle | Accuracy | Sensitivity | Specificity | Precision | Balanced Accuracy | F1 score | AUC |
|---|---|---|---|---|---|---|---|
| Logistic | 0,9582 | 0,0311 | 0,9982 | 0,4286 | 0,5147 | 0,0581 | 0,8387 |
| Lasso | 0,9584 | 0,0242 | 0,9987 | 0,4375 | 0,5114 | 0,0459 | 0,8385 |
| Ridge | 0,9582 | 0,0242 | 0,9985 | 0,4118 | 0,5114 | 0,0458 | 0,8382 |
| Elastic net | 0,9587 | 0,0138 | 0,9994 | 0,5000 | 0,5066 | 0,0269 | 0,8357 |
| KNN | 0,9549 | 0,0761 | 0,9928 | 0,3143 | 0,5345 | 0,1226 | 0,6828 |
| SVC | 0,9597 | 0,0450 | 0,9991 | 0,6842 | 0,5220 | 0,0844 | 0,7064 |
| Neural network | 0,9575 | 0,1176 | 0,9937 | 0,4474 | 0,5557 | 0,1863 | **0,8479** |
| CART | 0,9318 | 0,2595 | 0,9608 | 0,2219 | 0,6101 | 0,2392 | 0,6101 |
| Random forest | 0,9601 | 0,0761 | 0,9982 | 0,6471 | 0,5372 | 0,1362 | **0,8515** |
| XGB | 0,9602 | 0,0934 | 0,9976 | 0,6279 | 0,5455 | 0,1627 | **0,8585** |

TABLEAU 5.1 – Comparaison des résultats des classificateurs sur l'échantillonage d'évaluation

En général, il est constaté que les performances des modèles d'apprentissage automatique sont meilleures que celles des modèles naïfs d'où l'efficacité des traitements des données. Par ailleurs, on observe que les autres métriques d'évaluation sont complètement arbitraires et faux par rapport à la métrique principale AUC. Par exemple, l'exactitude (*Accuracy*) du SVC est très élevée, ce n'est néanmoins pas un bon classificateur quant à son AUC. Un autre exemple est la grande *Balanced Accuracy* du CART, étant donné que le CART est un modèle naïf qui produit essentiellement un problème de sur-apprentissage. Ainsi, on a désormais décidé de nous concentrer uniquement sur la métrique d'AUC et d'enlever toutes les autres.

Parmi la liste des modèles, les trois les plus performants sont le réseau de neurones, la forêt aléatoire et l'*Extreme Gradient boosting* (XGB) et ils atteignent un AUC moyen supérieur à 0,85. Ces modèles sont considérés comme les modèles complexes ou les méta-modèles. Il faut noter que le SVC est aussi un modèle avancé, mais ce classificateur ne fonctionne pas bien dans notre étude. D'ailleurs, la figure 5.1 expose la courbe ROC de ces trois meilleurs modèles, les AUCs sont bien présentés par la superficie sous la courbe. Nous observons que même si le résultat de la forêt aléatoire sur l'échantillon de test est bon, il provoque un grand problème de sur-apprentissage car sa prédiction sur le *train set* est parfaite (i.e. son ROC est en forme de l'angle). Il existe également le sur-apprentissage dans le XGB, pourtant cela est plus acceptable.

Les courbes ROC des trois meilleurs modèles sont affichées dans le graphique suivant :

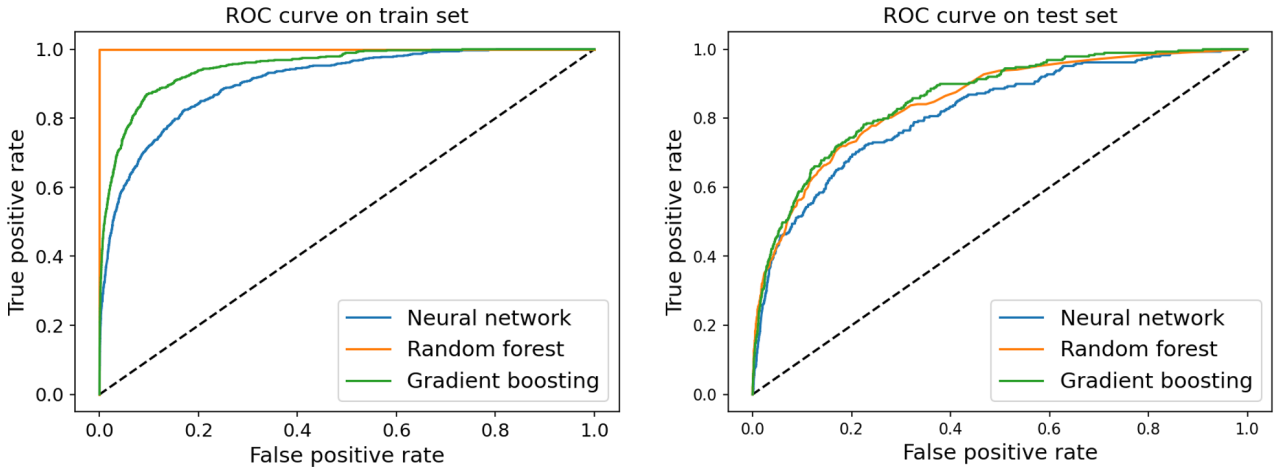



FIGURE 5.1 – Les courbes ROC des trois meilleurs classificateurs sur l'échantillon d'apprentissage (à gauche) et l'échantillon d'évaluation (à droite)

Puis, on fait retourner les algorithmes d'apprentissage combinés avec les différentes techniques d'échantillonage. Le tableau 5.2 suivant illustre le résultat de notre deuxième tentative. Dans la plupart des cas, la technique *Balanced (Bagging)* nous permet de renforcer les classificateurs, tandis que les sur- et sous-échantillonage ne sont pas efficace. Notamment, le *Blanced bagging CART* qui transforme le CART de modèle naïf en méta-modèle améliore largement son AUC.

| | **Logistic** | **Lasso** | **Ridge** | **Elastic net** | **KNN** | **SVC** | **Neural network** | **CART** | **Random forest** | **XGB** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Echantillon initial** | 0,8387 | 0,8385 | 0,8382 | 0,8357 | 0,6828 | 0,7064 | 0,8479 | 0,6101 | 0,8515 | **0,8585** |
| **Sur-échantillonage** | 0,8391 | 0,8408 | 0,8409 | 0,7553 | 0,6849 | 0,8192 | 0,7945 | 0,5777 | 0,8458 | 0,8443 |
| **Sous-échantillonage** | 0,8346 | 0,8401 | 0,8402 | 0,7267 | 0,7762 | 0,7904 | 0,7975 | 0,6911 | 0,8469 | 0,8369 |
| **Balanced (Bagging)** | 0,8403 | 0,8425 | 0,8423 | 0,8181 | 0,8112 | 0,7930 | 0,8467 | **0,8620** | **0.8640** | 0,8329 |

TABLEAU 5.2 – AUC (sur l'échantillon de test) des divers algorithmes d'apprentissage combinés avec les techniques de ré-échantillonage

Pour finaliser la modélisation, nous choisissons dans le tableau 5.2 les trois meilleures méthodes (XGB, Balanced CART et Balanced RF) pour construire le modèle de génération empilée. Le choix de plus que trois peut nous conduire à la computation intensive et l'inclusion des modèles moins performants pourrait créer un impact négatif sur l'estimateur de *Stacking*. La régression logistique a été utilisée comme le modèle de niveau 1 de *Stacking* puisque nous sommes en cas de classification.

Les courbes ROC et les AUC du modèle *Stacking* sont présentés respectivement sur la figure 5.2 et dans le tableau 5.3. L'AUC sur la base de test est de **0,8678** qui est la plus élevée en

comparant avec celles obtenues par les modèles de base. En agrégeant plusieurs estimateurs, le modèle *Stacking* pourrait réduire les erreurs et les limites des modèles de base, et pourrait également produire une meilleure estimation. La métrique d'évaluations du modèle de Stacking est résumée dans le tableau suivant :

| | **AUC** |
|---|---|
| **Echantillon d'apprentissage** | 0,9834 |
| **Echantillon d'évaluation** | **0,8678** |

TABLEAU 5.3 – Métriques d'évaluation du modèle *Stacking*

Les courbes ROC sur la base des données d'apprentissage et de test du modèle de Stacking seront présentées dans les graphiques suivants :

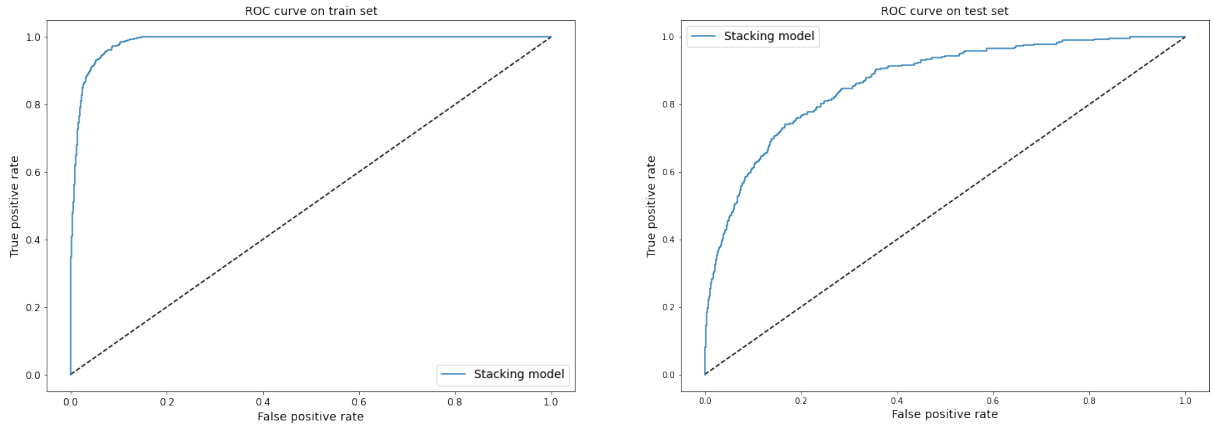



FIGURE 5.2 – Les courbes ROC du modèle de Stacking sur l'échantillon d'apprentissage (à gauche) et l'échantillon d'évaluation (à droite)

Enfin, nous avons décidé de retenir le modèle de Stacking comme le modèle final pour la prévision de turnover dans cette étude. Ensuite, le modèle a été utilisé pour estimer la probabilité de turnover sur toute la population de l'entreprise. Ce modèle donne évidemment une AUC élevée de **0,9826** sur la population totale.

Le tableau 5.4 suivant illustre les probabilités de turnover attendu par le modèle *Stacking* pour certains employées. Les salariés non-démissions ont des taux de turnover estimés très bas (mois de 0,06), inversement, les salariés ayant démissions ont des taux de turnover significativement élevés (de 0,15 à 0,71). Les résultats obtenus semblent cohérents et confirment la qualité de prévision et le choix du modèle final.

| | Resignation | Probability |
|---|---|---|
| 8007 | False | 0.053899 |
| 10834 | False | 0.055988 |
| 9294 | False | 0.010965 |
| 14932 | False | 0.007572 |
| 7147 | False | 0.009404 |
| 14155 | False | 0.004073 |
| 5504 | False | 0.015805 |
| 13452 | False | 0.004328 |
| 5485 | False | 0.004474 |
| 20578 | False | 0.033993 |

(A) Salariés sans turnover observé

| | Resignation | Probability |
|---|---|---|
| 8418 | True | 0.156224 |
| 11957 | True | 0.446622 |
| 11758 | True | 0.152292 |
| 16750 | True | 0.322829 |
| 6433 | True | 0.165460 |
| 16860 | True | 0.205307 |
| 16937 | True | 0.445921 |
| 10192 | True | 0.647212 |
| 3027 | True | 0.532836 |
| 19054 | True | 0.707791 |

(B) Salariés avec turnover observé

TABLEAU 5.4 – Probabilités de turnover

Le modèle final de Stacking a détecté les variables importantes comme dans le graphique suivant :

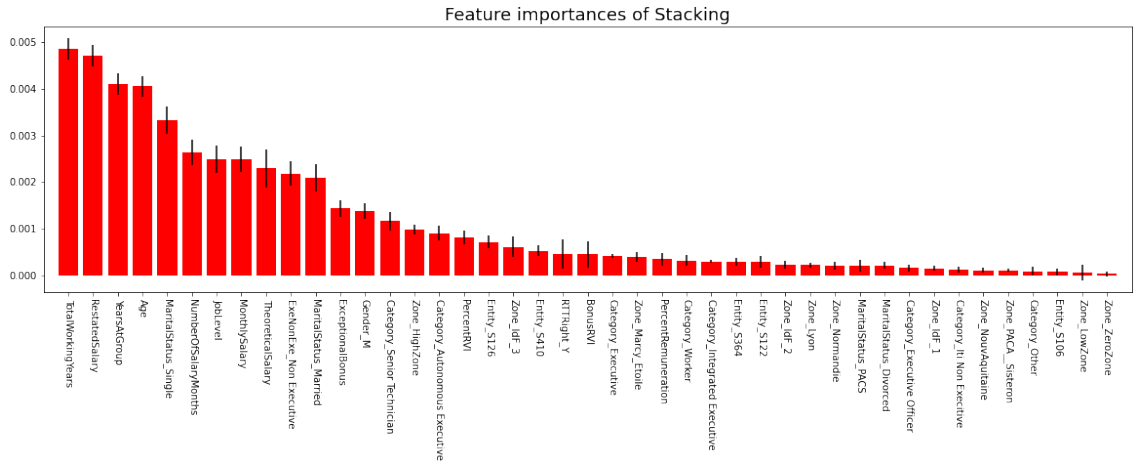



FIGURE 5.4 – Les plus importantes features détectées par le modèle *Stacking*

La figure 5.4 précise les prédicteurs ayant un impact important sur la prévision de turnover du modèle de Stacking, notamment les suivants : les éléments de salaire et de primes, l'ancienneté groupe, le statut matrimonial, le sexe et la zone. Cela confirme également quelques remarques constituées dans le chapitre 3.

### 5.1.2 Construction des nouvelles tables de turnover

Le but de ce mémoire ne se limite pas à estimer la probabilité de turnover dans un an, il vise à projeter dans le future et à construire une courbe de taux de turnover pour chaque salarié.

La projection demande plusieurs hypothèses et estimations sur à la fois les comportements sociaux et les comportements économiques. Plus précisément, les comportements sociaux qui pourront avoir des impacts sur le taux de turnover sont par exemple des changements de zone géographique, de fonction d'emploi (la catégorie), ou de statut marital, etc. Les comportements économiques pourraient être liés à la variation de l'inflation qui reflète la situation économique d'un pays, aux changements du secteur industriel de l'entreprise, ou aux changements des barèmes de rémunération, etc.

Dans cette étude, nous avons limité les hypothèses à retenir afin de réduire la complexité du modèle de projection, ainsi de rassurer la cohérence avec les hypothèses retenues dans le modèle de référence de l'évaluation des IFC.

Les hypothèses sont utilisées pour la projection comme les suivantes :

- Le sexe, le statut matrimonial et le droit RTT restent inchangés.
- Étant donné qu'il y a peu de transferts intra-groupe historiquement observés au sein de cette entreprise, on suppose que l'entité, la ville et la région du salarié restent inchangées.
- La catégorie est supposée inchangée pour être cohérente avec les hypothèses du modèle de référence.
- L'âge, l'ancienneté groupe et l'ancienneté carrière augmentent évidemment 1 chaque année de projection jusqu'à l'âge de départ à la retraite du salarié concerné.
- Le salaire annuel théorique, le salaire annuel reconstitué et le salaire mensuel augmentent 2,50 % par an pour les cadres et 2,00 % pour les non cadres qui sont en ligne avec les hypothèses du modèle de référence.
- Les autres éléments de primes et de nombre de mois de salaire sont supposés inchangés.

En combinant le modèle final de Stacking et les hypothèses de projection ci-dessus, une courbe de taux de turnover a été construite pour chaque salarié de l'entreprise, prise en compte les caractéristiques propres au salarié.

Les graphiques suivants montrent deux exemples des courbes de taux de turnover obtenues par la projection.

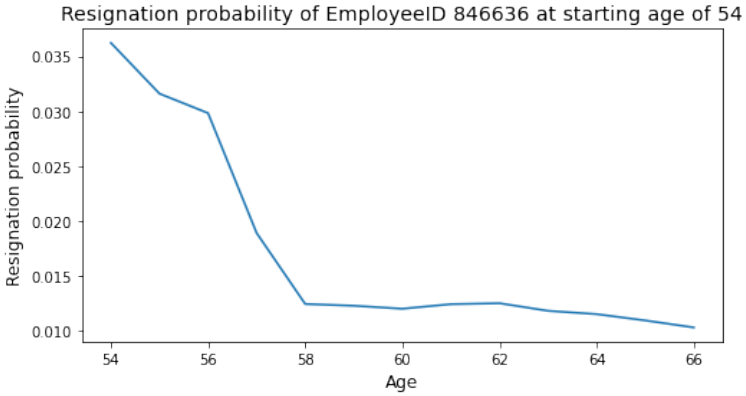



FIGURE 5.5 – Taux de turnover du salairé ID 846636


FIGURE 5.6 – Taux de turnover du salairé ID 198915

Le salarié ID 846636 dont l'âge actuel est de 54 ans a une courbe de taux de turnover avec les taux bas (moins de 4 %) et décroissants rapidement jusqu'à l'âge de 58 ans et puis décroissants graduellement vers 1 %) jusqu'à son âge de départ à la retraite de l'hypothèse (cf. Figure 5.5). Il est évidement et logique avec les remarques et les résultats obtenus que l'âge est négativement lié au turnover. Les salariés qui ont plus de 55 ans sont moins susceptibles de démissionner.

Contrairement, le salarié ID 198915 qui est plus jeune (33 ans actuellement) a une courbe des taux de turnover de manière différente (cf. Figure 5.6). Les taux de turnover ont une tendance de croissance jusqu'à l'âge de 42 ans avec le pic d'environ 30 %) et ensuite suivent une tendance à la baisse forte jusqu'à l'âge de 55 ans vers 2 %. A partir de 55 ans jusqu'à son âge de retraite, il suit une même tendance que celle du salarié ID 846636.

Il est important de noter que l'âge n'est pas qu'un seul facteur impactant le taux de turnover des salariés dans le modèle. La courbe des taux de turnover d'un salarié tient en compte tous les caractéristiques de ce salarié comme le sexe, le statut matrimonial, la catégorie, l'entité, les

salaires projetés, etc.

Dans cette étude, nous avons également construit une table de turnover prospective en agrégeant les courbes de taux de turnover individuelles. La table prospective obtenue tient compte des taux de turnover en fonction de l'âge et de l'année future (cf. **Annexe G**). Graphiquement, la table peut être présentée comme dans la figure 5.7 suivante.

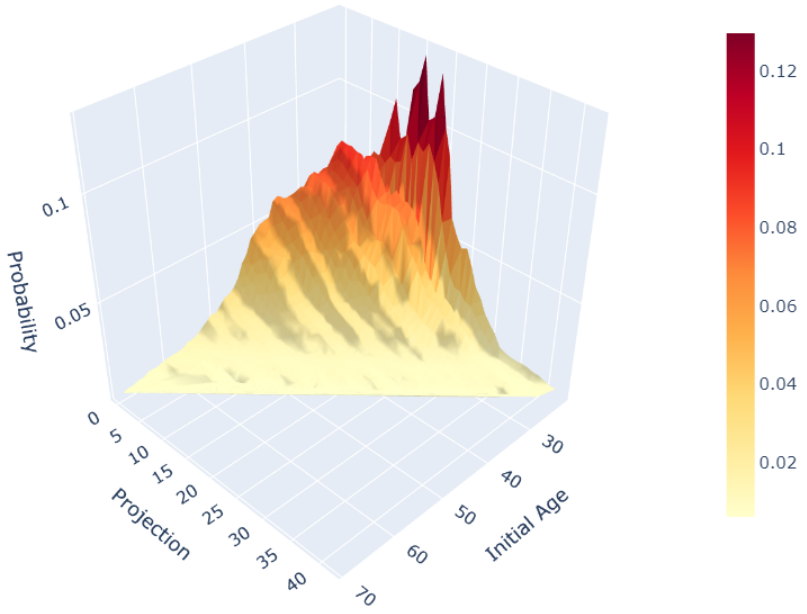



FIGURE 5.7 – Taux de turnover prospectifs

La structure des courbes de taux de turnover semble en ligne avec les observations et les analyses dans le chapitre 3, à l'exception d'un point de remarque comme suit :

- Les taux de turnover sont non nuls après 55 ans, bien que les taux ne soient que très faibles (contre les taux nuls après 55 ans dans les tables de turnover actuelles du modèle de référence). Cela s'explique par le fait que 3,60 % des 963 salariés ayant démissionné ont plus de 55 ans. Le modèle a appris et a reflété cette réalité observée en rajoutant des probabilités de turnover pour les années vers la retraite.

Les impacts des nouvelles tables sur l'engagement des IFC de l'entreprise seront présentés dans la section suivante.

## 5.2 Résultats actuariels par l'application des nouvelles tables de turnover

Cette section présentera les résultats de l'évaluation des IFC obtenus par l'application des nouvelles tables du turnover ainsi la reconnaissance comptable des impacts liés aux changements. Elle abordera ensuite la projection des engagements jusqu'à extinction de la population et proposera enfin le choix de la table de turnover à retenir.

### 5.2.1 Impacts des nouvelles tables sur l'engagement, la charge et les prestations futures

Les tables de turnover par individu et la table de turnover prospective obtenues dans la section précédente ont été appliquées dans le calcul des IFC au 31/12/2019 afin de mesurer les impacts de ces nouvelles tables. Pour la simplification, on appelle le « Modèle 1 » pour le modèle qui utilise les tables de turnover individuelles et le « Modèle 2 » pour le modèle qui utilise la table de turnover prospective agrégée pour l'ensemble des salariés de l'entreprise.

L'ensemble des calculs a été effectué sous Excel qui permet de réaliser les calculs de probabilité de présence de chaque salarié par son propre table de turnover ou les calculs par la table de turnover prospective. Les résultats avant et après application des nouvelles tables de turnover sont présentés dans le tableau suivant (montants en milliers d'euros) :

| **Modèles** | **Engagement DBO au 31/12/2019** | **Service Cost 2020** | **Interest Cost 2020** | **Charge 2020** |
|---|---|---|---|---|
| Modèle de référence | 959 706 | 55 981 | 7 113 | 63 094 |
| Modèle 1 | 821 440 | 43 818 | 6 077 | 49 984 |
| Écart en K€ | -138 307 | -12 163 | -1 036 | -13 199 |
| Écart en % | -14 % | -22 % | -15 % | -21 % |
| Modèle 2 | 799 709 | 43 181 | 5 913 | 49 095 |
| Écart en K€ | -159 997 | -12 799 | -1 200 | -13 999 |
| Écart en % | -17 % | -23 % | -17 % | -22 % |

Tableau 5.5 – Résultats avec l'application des nouvelles tables de turnover

Le graphique suivant permet d'illustrer ces résultats :

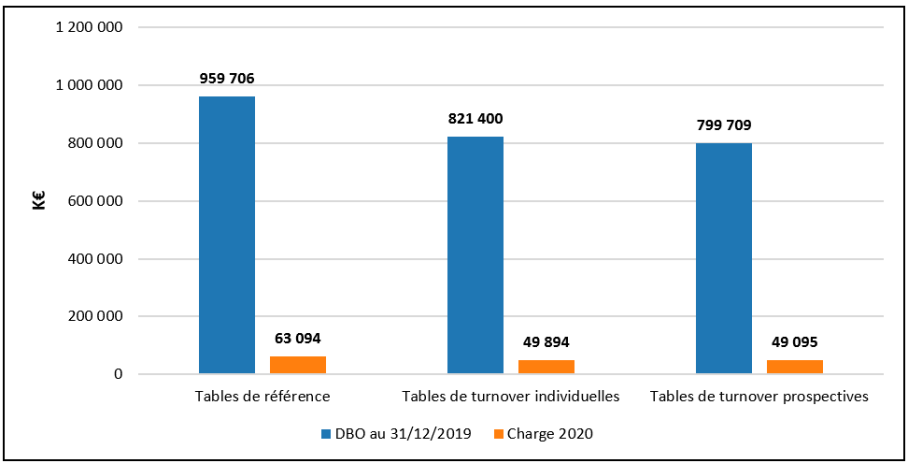



FIGURE 5.8 – Graphique des résultats avec l'application des nouvelles tables de turnover

Il est constaté que l'application des tables de turnover individuelles entraîne une baisse de l'ordre de 14 % sur l'engagement (DBO), soit d'environ 138 millions d'euros. Le coût des services et le coût d'intérêt étant positivement liés à l'engagement, la charge de l'exercice suivante a été diminuée 21 %, soit 13 millions d'euros en comparant avec celle du modèle de référence. La baisse de l'engagement s'explique par les raisons suivantes :

- Le modèle final de Stacking a prévu de nombreux salariés comme sortie en démission dont les taux de turnover sont significativement élevés en comparant avec ceux dans les tables de turnover du modèle de référence. Pour rappel, plus le taux de turnover est élevé, plus l'engagement est faible.
- Dans le modèle de référence, les taux de turnover sont en fonction de la catégorie et de l'entité d'un salarié, néanmoins, dans le modèle 1, à part de ces deux facteurs, il y a plusieurs d'autres facteurs ayant un impact important sur le turnover ont été intégrés. Or, les tables de turnover du modèle de référence ont été construites en 2016, avec certaines tables dont les taux de turnover sont considérablement faibles et fixes par tranche âge, les taux de turnover ne reflètent pas correctement les mouvements récents de la population. Par conséquent, les taux de turnover par salarié sont en général plus élevés que ceux dans les tables communes par entité.
- Les taux de turnover s'annulent à partir de 55 ans dans les tables du modèle de référence. Contrairement, comme expliqué dans la section précédente, la projection des taux de turnover par le modèle 1, ou également le modèle 2, tient en compte les probabilités de turnover au delà de 55 ans, bien que les taux ne soient que faibles. Les salariés ayant l'âge de 55 à 60 ans représentent 18,57 % de l'effectif, néanmoins leurs engagements représentent 31,60 % de l'engagement total de l'entreprise du modèle de référence (cf. Figure 2.3). L'engagement des salariés âgés contribue une partie importante car ces salariés sont proches à la retraite et nous avons utilisé la méthode avec prorata d'ancienneté pour

calculer l'engagement. Par conséquent, l'application les courbes de taux avec les taux non nuls après 55 ans (même s'ils sont très faibles) a diminué significativement l'engagement. Concernant les salariés plus de 60 ans, leur engagement total ne représente que mois de 5,00 % de l'engagement total, l'impact de baisse de l'engagement par les nouvelles courbes de taux de turnover existe, mais reste non matériel. (cf. Figure 5.9).

Les impacts du modèle 2 sont légèrement plus importants que ceux du modèle 1. Les raisons pour lesquelles ce modèle entraîne une baisse sur l'engagement en comparant avec le modèle de référence sont principalement comme les explications ci-dessus car la table de turnover prospective du modèle 2 est construite en agrégeant ou autrement dit en prenant la moyenne des tables individuelles du modèle 1.

Le graphique suivant présente la répartition de l'engagement par tranche d'âge de chaque modèle.

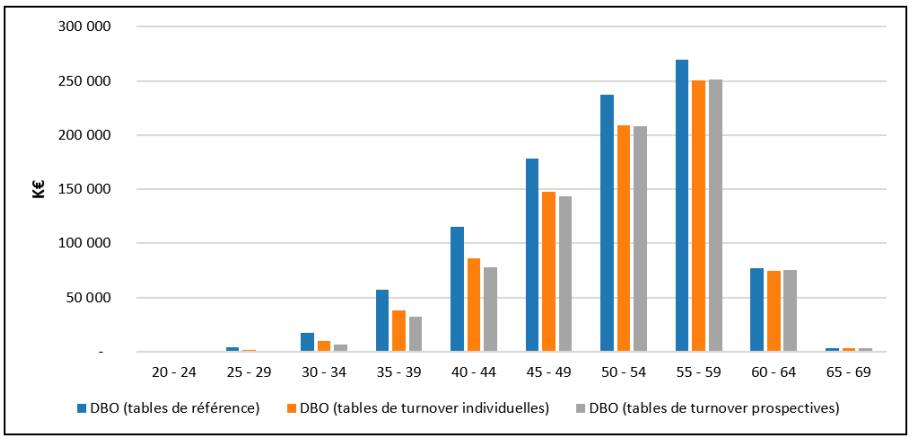



FIGURE 5.9 – Comparaison des engagements par tranche d'âges

Comme évoqué précédemment, les impacts sont importants pour les groupes des salariés de 50 à mois de 60 ans, et sont moins importants pour les groupes des salariés qui ont plus de 60 ans ou moins de 30 ans.

Il est également intéressant de vérifier les impacts des nouvelles tables sur les prestations attendues des IFC. Pour rappel, les prestations attendues sont les paiements probabilisés (sans actualisés et sans proratés) qui représentent les budgets futurs dont l'entreprise doit préparer à verser dans le futur.

Les prestations attendues sur les 10 prochaines années sont affichées dans le graphique suivant :

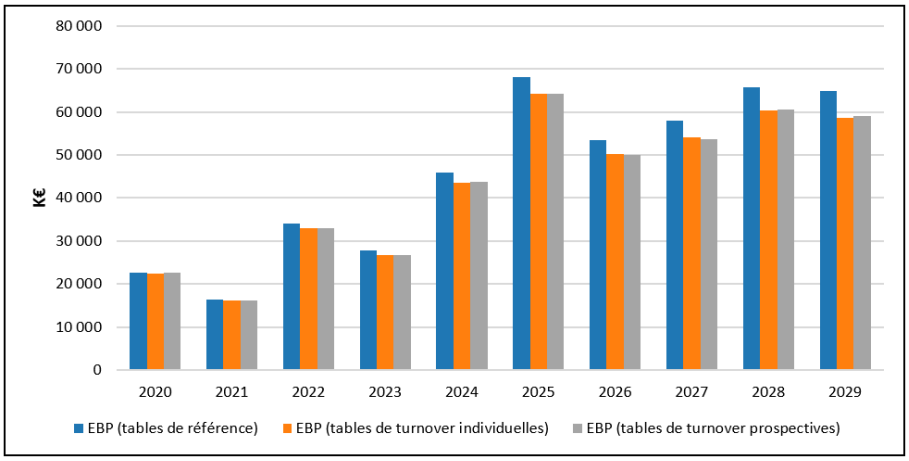



FIGURE 5.10 – Comparaison des prestations attendues

Le graphique montre que les impacts des nouvelles tables de turnover sur les prestations futures sont non-significatifs pour les 4 premières années. Cela est cohérent avec les observations précédentes sur les impacts négligeables des deux modèles 1 et 2 sur les salariés âgés plus de 60 ans qui bénéficieront de ces prestations attendues. Logiquement, les prestations des années à partir de 2024 des deux modèles ont significativement baissé par rapport avec celles du modèle de référence, car ces prestations seront versées pour les salariés de la tranche des âges de 55 à 59 ans.

Pour résumer, l'application des nouvelles tables de turnover entraîne les impacts importants sur l'engagement, la charge prévisionnelle et les prestations futures du régime des IFC de l'entreprise.

**Reconnaissance comptable des impacts liés aux changements des tables de turnover selon la norme IAS 19 :**

Il est important de noter l'engagement (DBO) au 31/12/2019 et la charge 2020 calculés par le modèle de référence ont été figés pour la clôture au 31/12/2019. A l'exception des changements importants dans le périmètre (acquisition, cession ou plans de départ volontaire, etc), ces montants seront donc déterminés comme les valeurs d'ouverture pour la clôture 2020 terminée au 31/12/2020. L'étude des impacts des nouvelles tables de turnover a été réalisée au 31/12/2019 avec la population 2019 afin de simuler les impacts du changement de cette hypothèse sur les résultats au 31/12/2020.

L'évolution de l'engagement au 31/12/2020 dépend de la population au 31/12/2020 et de la mise à jour des autres hypothèses actuarielles telles que la table mortalité et le taux d'actualisation. Sauf s'il y a des changements dans le périmètre, l'impact des nouvelles tables de

turnover sur l'engagement au 31/12/2020 sera du même ordre de grandeur avec les estimations au 31/12/2019 dans le tableau 5.5. Autrement dit, l'application des nouvelles tables de turnover du modèle 1 sur les évaluations actuarielles au 31/12/2020 avec la nouvelle population au 31/12/2020 entraînerait :

- une baisse de l'ordre de 14 % sur l'engagement DBO au 31/12/2020, soit d'environ 138 millions d'euros.
- une baisse de l'ordre de 21 % sur la charge prévisionnelle (la dotation) pour l'année 2021, soit d'environ 13 millions d'euros.

Selon la norme IAS 19 révisée, ces impacts du changement de l'hypothèse de turnover se traduisent :

- un impact de gain sur les capitaux propres (i.e. les autres éléments du résultats global - OCI) d'environ 138 millions d'euros. En effet, la provision des IFC au 31/12/2020 va baisser 138 millions d'euros.
- un impact sur le compte des résultats - P&L de l'année 2021 de baisse de la charge prévisionelle 2021 d'environ 13 millions d'euros.

Ainsi, ces impacts sont bénéfiques pour l'entreprise dans le sens où ils viennent baisser l'engagement et réduire la charge en compte des résultats.

### 5.2.2 Projection des résultats

Selon la norme IAS 19, le régime des IFC ne nécessite pas de projections de population et de résultats. Cependant, il est intéressant de projeter les engagements DBO jusqu'à extinction de la population pour voir l'évolution des engagements et pour comparer les différents scénarios dans le cas de notre étude. Dans cette étude, la population de l'entreprise partira complètement à la retraite en 2062.

Les engagements sociaux calculés conformément à la norme IAS 19 se projettent selon la formule suivante :

$$DBO_{31/12/n+1} = DBO_{31/12/n} + SC_{n+1} + IC_{n+1} - EBP_{n+1} \tag{5.1}$$

avec :

- $SC_{n+1}$ est le coût des services rendu pour l'exercice suivant (*Service Cost*) qui correspond à l'accélération probable de l'engagement due à une année supplémentaire de l'ancienneté du bénéficiaire.

- $IC_{n+1}$ est le coût d'intérêt de l'année suivante (*Interest Cost*) correspondant à l'augmentation de l'engagement liée à la réduction d'une année dans la période d'actualisation des prestations futures.
- $EBP_{n+1}$ sont les prestations probabilisées attendues (*Expected Benefit Payments*) à verser dans l'année qui diminuent l'engagement en fin d'année.

La projection des engagements se base sur l'hypothèse du taux d'actualisation de la clôture actuelle, c'est-à-dire dans notre étude, le taux d'actualisation de 0,75 % au 31/12/2019.

Les résultats des projections sont affichés dans le graphique suivant :

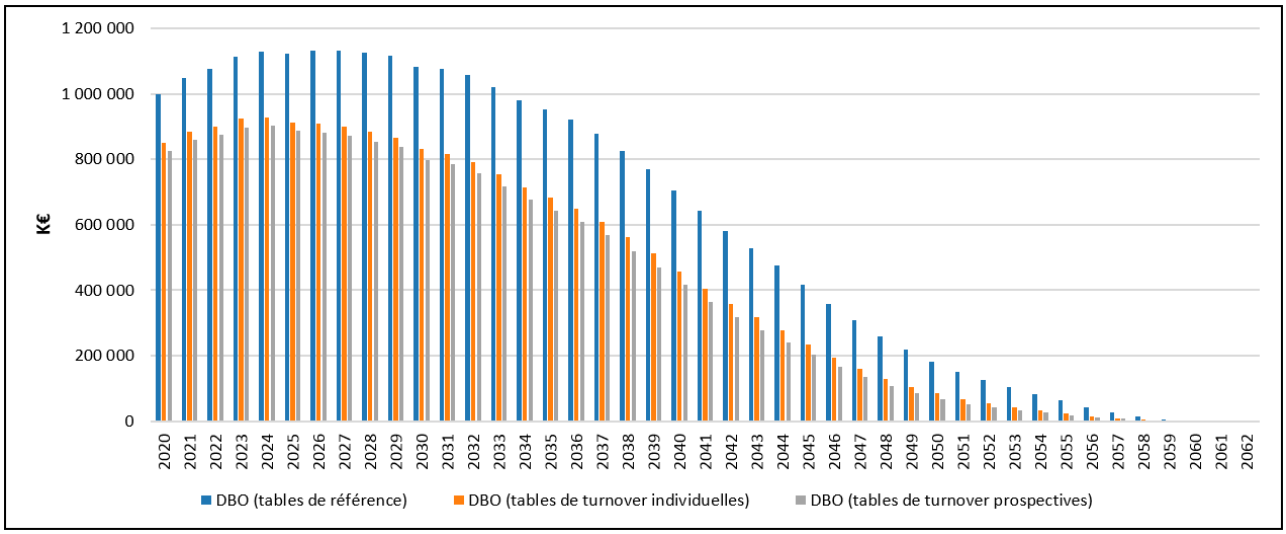



FIGURE 5.11 – Comparaison des engagements projetés

Le graphique montre que la majorité des engagements se situent entre les années de 2024 à 2034 où la part la plus important de population partira à la retraite. Cela est en ligne avec les analyses précédentes. En effet, la population de l'entreprise est relativement âgée avec un âge moyen de 46,22 ans et les salariés de 50 ans à 60 ans représentent un tiers de la population, ce qui signifie que ces personnes partiront massivement à la retraite dans 5 à 15 ans, i.e. dans les années de 2024 à 2034. Le graphique illustre également les impacts importants des nouvelles tables de turnover du modèle 1 et du modèle 2 sur les engagements projetés.

Les écarts des résultats entre le modèle 1 et le modèle 2 étant non significatifs, le choix de la table de turnover à retenir dépend le préférence de l'actuaire et de l'entreprise au regard de la complexité des modèles de calculs. Les tables de turnover individuelles donnent les meilleurs estimations des probabilités de turnover pour chaque salarié, néanmoins elles demandent des calculs considérablement lourds dans le modèle 1. Contrairement, la table de turnover prospec-

tive du modèle 2 est une moyenne ou une « proxy » des tables individuelles du modèle 1, ce qui pourrait conduire une légère sur-estimation du turnover dont l'impact sur l'engagement est un peu plus important. Cependant, le modèle d'évaluations des IFC utilisant la table de turnover prospective ne nécessite pas d'une volume énorme de calculs.

Ainsi, la table de turnover prospective du modèle 2 pourrait être proposée à l'entreprise.

## 5.3 Limitations

Cette dernière section a pour but de présenter et de résumer les limites auxquelles nous nous confrontons dans l'étude. Ce sont les suivantes :

**Sur les données :**

- La base des données contient une vingtaine de variables explicatives, néanmoins la diversité de ce type d'information est limitée. Plusieurs prédicteurs qui pourraient avoir un impact sur le turnover n'ont pas été pris en compte dans cette étude comme le niveau de formation, les enfants à charge, la satisfaction d'emploi, la distance entre le lieu habituel de travail et le domicile, etc. Cependant, ces données ne sont pas disponibles dans l'entreprise ou nécessitent un travail supplémentaire intensif pour être collectées. Par ailleurs, il faut noter que l'ajout des variables ne promet pas toujours une amélioration de la qualité de modèles de prévision. Ce sont les raisons pour lesquelles nous n'avons pas essayé d'intégrer les autres informations dans les données.
- Il existe plusieurs techniques ou des possibilités de traitement des variables explicatives. Nous en avons essayé certaines, néanmoins les performances des modèles n'ont pas été significativement augmentées. Par exemple, nous avons lancé la méthode de l'analyse en composantes principales (ACP) afin de résoudre le problème de fortes corrélations et de multicolinéarité. Cette technique conduit à la computation intensive et ne rend pas toujours les meilleurs résultats (cf le tableau 5.6 suivant). D'ailleurs, dans le but d'examiner l'efficacité des techniques de ré-échantillonage, nous avons gardé la base des données initiale et n'avons pas donc pris en compte des autres techniques de traitement de données.

| | **Logistic** | **Lasso** | **Ridge** | **Elastic net** | **KNN** | **SVC** | **Neural network** | **CART** | **Random forest** | **XGB** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Echantillon initial** | 0,8387 | 0,8385 | 0,8382 | 0,8357 | 0,6828 | 0,7064 | 0,8479 | 0,6101 | 0,8515 | 0,8585 |
| **Echantillon avec l'ACP** | 0,8342 | 0,8368 | 0,8366 | 0,6399 | 0,6893 | 0,6357 | 0,8226 | 0,6073 | 0,8415 | 0,8555 |

TABLEAU 5.6 – AUC (sur l'échantillon de test) des divers algorithmes d'apprentissage combinés avec la technique de l'ACP

**Sur les algorithmes d'apprentissage automatique :** Certains des algorithmes de machine learning utilisés dans cette étude prennent des hyper-paramètres à sélectionner et à optimiser. Par exemple, concernant le modèle SVC, nous avons utilisé la fonction de noyau de base radiale. Cependant, il existe plusieurs types de fonction de noyau qui n'ont pas été challengés dans notre étude tels que la polynomiale, le réseau de neurones, etc. Il y a également plusieurs modèles de classification qui n'ont pas été utilisés dans cette étude comme les suivants : l'analyse discriminante linéaire, la classification naïve bayésienne, etc. Néanmoins, les performances d'un modèle prédictif dépendent généralement plus de la nature et la qualité des données, du soin apporté à leur préparation et à leur sélection, que de la technique de modélisation elle-même (Tuffery, S., 2005). En effet, nous n'avons pas essayé de lancer des autres modèles ou des hyper-paramètres en raison du compromis entre une amélioration marginale des résultats et le temps et les efforts consacrés.

**Sur les hypothèses de projection :** Comme discuté dans la partie 5.1.2, la construction des nouvelles tables de turnover ne tient pas compte de l'évolution de plusieurs facteurs ou de comportements futurs. Hormis des exemples cités dans la 5.1.2, l'évolution des taux de turnover pourrait être impactée par d'autres événements comme la crise sanitaire. Il est évident que la pandémie Covid 19 a des conséquences importantes à la fois pour l'économie et pour la société. La décision de démissionner et le mouvement des salariés font partie de ces conséquences. Par exemple, dans certains secteurs, le taux de turnover a diminué en 2020 en raison de la situation incertaine. A l'inverse, dans d'autres secteurs (par exemple le conseil), les salariés ont été tentés de quitter leur entreprise ou leur région. Le taux de rotation du personnel de l'entreprise augmenterait en conséquence. Étant donnée que ces facteurs nécessitent des modèles de prévision assez complexes, nous n'avons pas tenu compte de ces facteurs dans nos hypothèses de projection et les avons considérés comme l'orientation de recherches futures.

# Conclusion générale

Dans cette étude, l'importance de l'hypothèse de turnover a été soulignée dans le calcul des engagements sociaux, notamment pour les Indemnités de Fin de Carrière. En effet, IAS 19 propose des méthodes de calcul et des hypothèses actuarielles de plus en plus précises sur les IFC, néanmoins le taux de turnover reste encore une hypothèse peu régulée par la norme. Cependant, dans le contexte économique actuel, où les taux d'actualisation sont historiquement bas, et les hypothèses démographiques autres que le taux de turnover varient peu, cette dernière joue un rôle crucial. C'est une hypothèse essentielle à établir et à revoir régulièrement car elle varie d'une entreprise à l'autre et dépend de plusieurs paramètres.

Les tables de turnover actuelles de l'entreprise ont été construites en 2016 et sont en fonction de l'âge, de l'entité et de la catégorie. Or, plusieurs facteurs pourraient avoir des impacts importants sur le turnover d'un salarié n'ont pas été pris en compte. En effet, ce mémoire a utilisé de nombreuses techniques d'apprentissage automatique pour estimer le taux de turnover et a proposé des nouvelles tables basées sur l'ensemble des données de ressources humaines des salariés.

Les algorithmes de machine learning qui ont été examinés dans cette étude sont : les modèles linéaires (régression logistique et ses versions pénalisées), les modèles non-linéaires (K plus proches voisins, Machines à vecteurs de support et Réseau de neurones) et quelques modèles basés sur les arbres de décision (CART, Forêts aléatoires, Extreme Gradient Boosting (XGB) et modèle de Stacking). D'ailleurs, les données de l'étude sont fortement déséquilibrées car le taux de turnover moyen observé représente seulement 4,13 % de la population. Cela a impacté la qualité de prévision de plusieurs modèles. Par conséquent, ce mémoire a abordé plusieurs techniques afin de résoudre ce problème telles que celles de ré-échantillonage ou encore de pondération de la fonction de perte. Parmi ces méthodes, en se basant sur les AUC obtenues, les modèles les plus performants dans la prévision du turnover sont : Réseau de neurones, XGB, *Balanced CART* et *Balanced Random Forest*. Enfin, en appliquant le modèle de généralisation empilée (*Stacking*), le modèle final a été calibré avec la meilleure estimation dans cette étude.

Les modèles de machine learning ont détecté plusieurs facteurs ayant un impact significatif sur le turnover, notamment : les éléments de salaire et de bonus, l'ancienneté, l'âge et le statut matrimonial. Il faut noter qu'il existe plusieurs d'autres variables qui pourraient affiner le turnover mais elles ne sont pas étudiées dans ce mémoire comme par exemple les facteurs propres au salarié (niveau de formation, satisfaction d'emploi, etc.) et les facteurs extérieurs à l'entreprise (conjoncture économique, changements dans le secteur industriel, changements juridiques, etc). Étant donné qu'il est difficile et complexe de collecter ces données et de construire les modèles

basés sur ces facteurs, ils ne sont pas intégrés dans cette étude. Néanmoins, cela pourrait être des futurs axes de recherche.

Basé sur le modèle d'apprentissage automatique retenu, les hypothèses de projection ont été proposées afin de construire des tables de turnover individuelles ainsi que d'agréger une table prospective pour l'ensemble des salariés. Dans le but de réduire la complexité du modèle de projection et de respecter les hypothèses du modèle de référence des calculs des IFC, quelques facteurs ont été proposés inchangés et les autres, indiqués dans le paragraphe précédent, n'ont pas été pris en compte dans la projection. Concernant les nouvelles tables de turnover, elles répondent bien aux analyses et statistiques observées, et s'adaptent aux exigences de la norme IAS 19. Elles pourraient alors être proposées au client lors de la clôture 2020.

Ce mémoire a aussi montré le fait que l'hypothèse de turnover a un impact important sur l'engagement et donc sur la provision des passifs sociaux de l'entreprise. L'application des nouvelles tables de turnover entraînerait une baisse significative sur l'engagement de l'ordre de 14 % soit environ 138 millions d'euros pour le modèle 1 (tables individuelles), et de l'ordre de 17 % soit environ 159 millions d'euros pour le modèle 2 (table prospective). Les impacts sur la charge de l'exercice suivant sont également importants : une baisse d'environ 22 %, soit environ 14 millions d'euros. Ces impacts sont bénéfiques pour l'entreprise car ils viennent baisser l'engagement et réduire la charge en compte des résultats.

Pour conclure, ce mémoire a traité de la problématique de l'hypothèse de taux de turnover en utilisant des techniques de machine learning dans le cadre des évaluations des IFC en norme IAS 19. Cela a également montré les opportunités et les avantages de machine learning qui pourraient aider à estimer le taux de turnover d'un salarié et à construire des nouvelles tables. Les techniques d'apprentissage automatique pourront améliorer la qualité de prévision avec les évolutions des données et le développement de nouvelles techniques. Cependant, comme le turnover peut fortement faire varier l'engagement, c'est un point auquel les actuaires doivent prêter une attention toute particulière dans le calcul des engagements sociaux afin de se conformer à la norme IAS 19.

# Bibliographie


Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*. Series B (Methodological, 26(2), 211–252. Récupéré de : https://doi.org/10.2307/2287791

DREES, Les retraités et les retraites, édition 2020. Récupéré de : https://drees.solidarites-sante.gouv.fr/publications-documents-de-reference/panoramas-de-la-drees/les-retraites-et-les-retraites-edition

Ecary, B., RETRAITE - Support de cours 3, ISFA, Année 2013-2014.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters. *Pattern Recognition Letters*, 27, 861–874. Récupéré de : http://www.sciencedirect.com/science/article/pii/S016786550500303X

Focus IFRS, *IAS 19 Avantages du personnel (version 2013)*. Date de mise à jour : 02/04/2019. Récupéré de : http://www.focusifrs.com/menu_gauche/normes_et_interpretations/textes_des_normes_et_interpretations/ias_19_avantages_du_personnel_version_2013

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive Logistic Regression : A Statistical View of Boosting : Discussion. *The Annals of Statistics*, 28(2), 374–377. Récupéré de : http://projecteuclid.org/euclid.aos/1016218223

Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. *MIT Press*. Récupéré de : http://www.deeplearningbook.org

Greene, William H. (2008). Econometric Analysis, 6th Edition, Upper Saddle Rive, NJ : Prentice-Hall. Récupéré de : https://spu.fem.uniag.sk/cvicenia/ksov/obtulovic/Mana%C5%BE.%20%C5%A1tatistika%20a%20ekonometria/EconometricsGREENE.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning (2nd edition)*. Récupéré de : https://doi.org/10.1007/978-0-387-84858-7

International Accounting Standards Board (2011). IAS 19 Employee Benefits.

Kuhn, M., & Johnson, K. (2016). *Applied Predictive Modeling* (5th ed.). New York : Springer. Récupéré de : https://doi.org/10.1007/978-1-4614-6849-3

La retraite en clair, *L'épargne retraite : les solutions existantes*. Date de mise à jour : 29/07/2020. Récupéré de : https://www.la-retraite-en-clair.fr/cid3190641/l-epargne-retraite-les-solutions-existantes.html

Milhaud, X. (2018). Data science en actuariat - Support de cours, Master 2 Actuariat, ISFA.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. *NIPS-W*. Récupéré de : https://openreview.net/pdf?id=BJJsrmfCZ

Shmueli, G., Patel, N., & Bruce, P. (2010). Data mining for business intelligence. Hoboken. Récupéré de : https://scholar.google.com/scholar?hl=nl&q=hmueli%2C+G.%2C+Patel%2C+N.+R.%2C+%26+Bruce%2C+P.+C.+%282010%29.+Data+mining+for+business+intelligence.+Hoboken.&btnG=&lr=

Tuffery, S. (2005), Améliorer les performances d'un modèle prédictif : perspectives et réalité. *DMAS*, vol. RNTI-A-1, 45-72. Récupéré de : https://editions-rnti.fr/render_pdf.php?p=1001487&p1

Witten, I., Frank, E., Hall, M., & Pal, C. (2016). Data Mining : Practical machine learning tools and techniques. Récupéré de : https://books.google.com/books?hl=nl&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=Witten,+I.+H.,+Frank,+E.,+Hall,+M.+A.,+%26+Pal,+C.+J.+(2016).+Data+Mining:+Practical+machine+learning+tools+and+techniques.+Morgan+Kaufmann.&ots=8HIJycgEx7&sig=2tUWfBEV2a-F3ycLsC4WP

Wolpert, David H. (1992). Stacked generalization. *Neural Networks*, Volume 5, Issue 2, Pages 241-259. Récupéré de : https://www.sciencedirect.com/science/article/abs/pii/S0893608005800231

# Annexes

## Annexe A : Récapitulatif des caisses de retraites par régimes

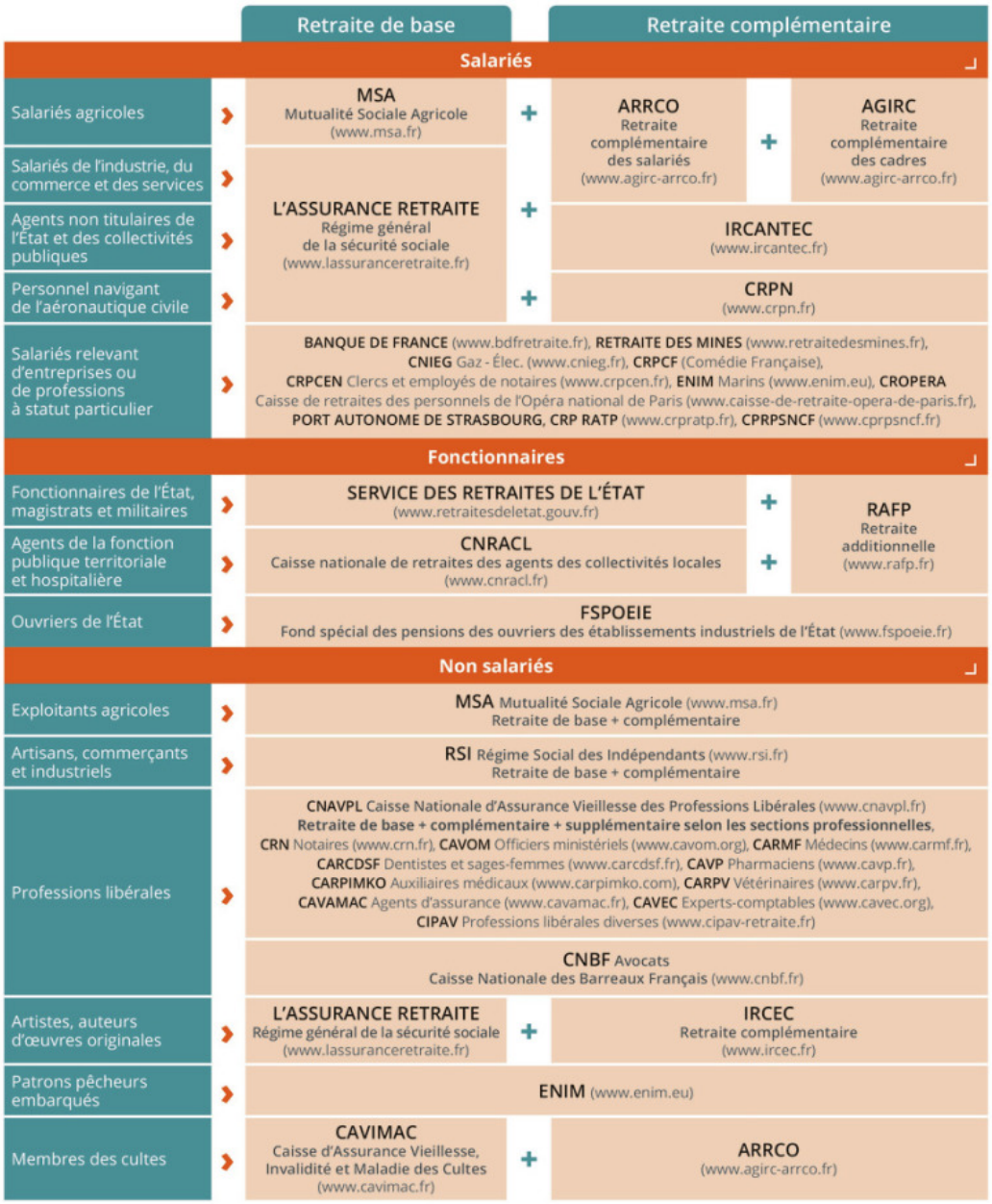



FIGURE 5.12 – Les différents régimes de retraite en France

# Annexe B : Extraits de la note de la CNCC (2018)

ENGAGEMENTS DE RETRAITE ET AVANTAGES SIMILAIRES – Application de la recommandation n° 2013-02 de l'Autorité des normes comptables – IFRS – Détermination du taux de rotation du personnel pour les besoins de l'évaluation de l'engagement ou de la provision – Prise en compte des prévisions de démission (oui) – Prise en compte des licenciements et des ruptures conventionnelles (non).

(EC 2018-17)

Une entreprise de plus de 250 salariés applique la recommandation n° 2013-02 de l'Autorité des normes comptables du 7 novembre 2013 relative aux règles d'évaluation et de comptabilisation des engagements de retraite et avantages similaires pour les comptes annuels et les comptes consolidés établis selon les normes comptables françaises.

Le seul engagement de retraite supporté par l'entreprise est constitué par les indemnités de fin de carrière du personnel, la législation française prévoyant que des indemnités sont versées aux salariés au moment de leur départ en retraite, en fonction de leur ancienneté et de leur salaire à l'âge de la retraite.

Que l'entreprise provisionne ses engagements retraite ou qu'elle les mentionne dans l'annexe de ses comptes, et qu'elle ait adopté la méthode 1 ou la méthode 2 de la recommandation ANC n° 2013-02, il convient de déterminer le taux de rotation du personnel utilisé dans le cadre de l'évaluation des indemnités de fin de carrière (IFC).

[...]

**Réponse de la Commission commune de doctrine comptable**

*Référentiel comptable français*

Le droit comptable français n'impose pas la comptabilisation au passif des engagements de retraite et assimilés et offre la possibilité de mentionner ces engagements dans l'annexe. Il ne contient aucune disposition prescriptive sur la manière d'évaluer les engagements, que l'entité ait choisi de les provisionner ou de les mentionner en annexe.

Dans ce contexte, l'Autorité des normes comptables a publié une recommandation citée ci-dessus qui fournit des dispositions sur le mode d'évaluation des engagements. La question soulevée se situe uniquement dans le cadre de l'application de cette recommandation et provient de ce que

celle-ci n'indique pas comment déterminer le taux de rotation du personnel pour les besoins de l'évaluation de l'engagement ou de la provision. La réponse donnée ci-dessous à la question est fournie dans le cadre de la recommandation de l'Autorité des normes comptables et ne préjuge pas de celles qui seraient données si l'évaluation de l'engagement ou de la provision était effectuée selon d'autres méthodes que celles préconisées par la recommandation.

Une entité est tenue de verser des indemnités de fin de carrière à ses salariés, sous condition que ceux-ci soient encore en activité au sein de l'entité au moment où ils liquident leur droit à retraite. Si l'entité choisit de licencier un salarié avant son départ à la retraite, ou si elle convient avec lui d'une rupture anticipée de son contrat de travail, elle sera tenue de lui verser une indemnité. Quoique les faits générateurs de l'indemnité de fin de carrière et de l'indemnité de licenciement soient différents, leur montant est généralement calculé sur les mêmes bases (en pourcentage de mois de salaire par année d'ancienneté), l'indemnité de licenciement due étant pratiquement toujours supérieure à l'indemnité de fin de carrière. Ainsi, qu'il s'agisse du départ à la retraite, du licenciement ou de la rupture conventionnelle, l'entité est tenue de payer une indemnité au salarié qui la quitte. Le seul cas de départ du salarié n'engendrant le paiement d'aucune indemnité est celui de la démission.

Dans ce contexte, la Commission est d'avis que l'évaluation des indemnités de fin de carrière doit être effectuée en tenant compte des seules prévisions de démission, dans la mesure où tout autre cas de départ avant l'âge de la retraite engendre pour l'entité un paiement au moins aussi important que l'indemnité de fin de carrière.

L'indemnité de fin de carrière constitue le minimum qui sera payé en tout état de cause, hormis le cas des démissions. La Commission considère que le taux de rotation à prendre en considération dans les hypothèses démographiques prévues par le paragraphe 6232 de la recommandation de l'Autorité des normes comptables doit en conséquence être déterminé en ne tenant compte que des seules prévisions de démission, à l'exclusion de toute autre hypothèse de départ.

*Normes IFRS*

Le raisonnement développé selon les normes IFRS pour répondre à la question soulevée est sensiblement identique à celui décrit ci-dessus selon les règles comptables françaises.

Dans le contexte français, la loi prévoit le versement d'une indemnité légale ou conventionnelle minimum en cas de licenciement ou rupture conventionnelle, qui est en général d'un montant proche, voire supérieur, à celui de l'indemnité de départ à la retraite qui aurait été obtenue si l'employé était resté au service de l'entreprise jusqu'à ce qu'il puisse en bénéficier.

En procédant à des licenciements, l'entreprise s'exonère de son obligation de payer des indem-

nités de départ à la retraite, mais elle ne peut le faire qu'en lui substituant une autre obligation, en général plus onéreuse, celle de payer des indemnités de licenciements. Or, les licenciements étant sous le contrôle de l'entreprise, ils ne peuvent être provisionnés qu'à la date d'annonce des licenciements conformément aux conditions imposées par IAS 19.165-167.

En conséquence, tenir compte des futurs licenciements dans le calcul du taux de rotation retenu pour calculer l'engagement de retraite aboutirait à sous-évaluer les provisions reconnues au bilan au titre des indemnités de départ à la retraite. Au regard d'IAS 19, quelle que soit la cause de départ du salarié, hormis le cas d'une démission, qu'il s'agisse d'un départ en retraite ou d'un licenciement, la société est tenue de verser une indemnité minimum, de manière certaine. Seule la date du versement est incertaine. Les formes juridiques d'indemnisation (indemnité de licenciement ou indemnité de départ à la retraite) sont mutuellement exclusives mais l'indemnité de départ à la retraite peut être considérée comme un montant minimum certain. Or le paragraphe 164 de la norme IAS 19 impose de provisionner les indemnités dont le paiement est certain en tant qu'avantage postérieur à l'emploi et non en tant qu'indemnité de rupture du contrat de travail.

En résumé, compte tenu du contexte français particulier, il n'est pas possible de raisonner in abstracto au niveau de la seule provision pour indemnité de départ à la retraite.

# Annexe C : Tables de l'âge de départ à la retrait et de l'âge de début de carrière

| Année de naissance | Age retraite minimum taux plein | Nombre de Trimestres nécessaires | Age annulation décote Cadres | Age annulation décote Non Cadres |
|---|---|---|---|---|
| 1936 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1937 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1938 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1939 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1940 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1941 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1942 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1943 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1944 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1945 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1946 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1947 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1948 | 60,00 | 160,00 | 63,00 | 60,00 |
| 1949 | 60,25 | 161,00 | 63,25 | 60,25 |
| 1950 | 60,50 | 162,00 | 63,50 | 60,50 |
| 1951 | 60,75 | 163,00 | 63,75 | 60,75 |
| 1952 | 61,00 | 164,00 | 64,00 | 61,00 |
| 1953 | 61,25 | 165,00 | 64,25 | 61,25 |
| 1954 | 61,58 | 165,00 | 64,25 | 61,58 |
| 1955 | 62,00 | 166,00 | 64,50 | 62,00 |
| 1956 | 62,00 | 166,00 | 64,50 | 62,00 |
| 1957 | 62,00 | 166,00 | 64,75 | 62,00 |
| 1958 | 62,00 | 167,00 | 65,25 | 62,25 |
| 1959 | 62,00 | 167,00 | 65,25 | 62,25 |
| 1960 | 62,00 | 167,00 | 65,50 | 62,50 |
| 1961 | 62,00 | 168,00 | 65,75 | 62,75 |
| 1962 | 62,00 | 168,00 | 66,00 | 63,00 |
| 1963 | 62,00 | 168,00 | 66,00 | 63,00 |
| 1964 | 62,00 | 169,00 | 66,50 | 63,50 |
| 1965 | 62,00 | 169,00 | 66,50 | 63,50 |
| 1966 | 62,00 | 169,00 | 66,75 | 63,75 |
| 1967 | 62,00 | 170,00 | 67,00 | 64,00 |
| 1968 | 62,00 | 170,00 | 67,00 | 64,25 |
| 1969 | 62,00 | 170,00 | 67,00 | 64,25 |
| 1970 | 62,00 | 171,00 | 67,00 | 64,75 |
| 1971 | 62,00 | 171,00 | 67,00 | 64,75 |
| 1972 | 62,00 | 171,00 | 67,00 | 64,75 |
| 1973 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1974 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1975 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1976 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1977 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1978 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1979 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1980 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1981 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1982 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1983 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1984 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1985 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1986 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1987 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1988 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1989 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1990 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1991 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1992 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1993 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1994 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1995 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1996 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1997 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1998 | 62,00 | 172,00 | 67,00 | 65,00 |
| 1999 | 62,00 | 172,00 | 67,00 | 65,00 |
| 2000 | 62,00 | 172,00 | 67,00 | 65,00 |
| 2001 | 62,00 | 172,00 | 67,00 | 65,00 |
| 2002 | 62,00 | 172,00 | 67,00 | 65,00 |
| 2003 | 62,00 | 172,00 | 67,00 | 65,00 |
| 2004 | 62,00 | 172,00 | 67,00 | 65,00 |
| 2005 | 62,00 | 172,00 | 67,00 | 65,00 |

TABLEAU 5.7 – Âge retraite

| Année de naissance | Cadres | Non Cadres |
|---|---|---|
| 1936 | 23,00 | 20,00 |
| 1937 | 23,00 | 20,00 |
| 1938 | 23,00 | 20,00 |
| 1939 | 23,00 | 20,00 |
| 1940 | 23,00 | 20,00 |
| 1941 | 23,00 | 20,00 |
| 1942 | 23,00 | 20,00 |
| 1943 | 23,00 | 20,00 |
| 1944 | 23,00 | 20,00 |
| 1945 | 23,00 | 20,00 |
| 1946 | 23,00 | 20,00 |
| 1947 | 23,00 | 20,00 |
| 1948 | 23,00 | 20,00 |
| 1949 | 23,00 | 20,00 |
| 1950 | 23,00 | 20,00 |
| 1951 | 23,00 | 20,00 |
| 1952 | 23,00 | 20,00 |
| 1953 | 23,00 | 20,00 |
| 1954 | 23,00 | 20,00 |
| 1955 | 23,00 | 20,00 |
| 1956 | 23,00 | 20,00 |
| 1957 | 23,25 | 20,25 |
| 1958 | 23,50 | 20,50 |
| 1959 | 23,50 | 20,50 |
| 1960 | 23,75 | 20,75 |
| 1961 | 23,75 | 20,75 |
| 1962 | 24,00 | 21,00 |
| 1963 | 24,00 | 21,00 |
| 1964 | 24,25 | 21,25 |
| 1965 | 24,25 | 21,25 |
| 1966 | 24,50 | 21,50 |
| 1967 | 24,50 | 21,50 |
| 1968 | 24,75 | 21,75 |
| 1969 | 24,75 | 21,75 |
| 1970 | 25,00 | 22,00 |
| 1971 | 25,00 | 22,00 |
| 1972 | 25,00 | 22,00 |
| 1973 | 25,00 | 22,00 |
| 1974 | 25,00 | 22,00 |
| 1975 | 25,00 | 22,00 |
| 1976 | 25,00 | 22,00 |
| 1977 | 25,00 | 22,00 |
| 1978 | 25,00 | 22,00 |
| 1979 | 25,00 | 22,00 |
| 1980 | 25,00 | 22,00 |
| 1981 | 25,00 | 22,00 |
| 1982 | 25,00 | 22,00 |
| 1983 | 25,00 | 22,00 |
| 1984 | 25,00 | 22,00 |
| 1985 | 25,00 | 22,00 |
| 1986 | 25,00 | 22,00 |
| 1987 | 25,00 | 22,00 |
| 1988 | 25,00 | 22,00 |
| 1989 | 25,00 | 22,00 |
| 1990 | 25,00 | 22,00 |
| 1991 | 25,00 | 22,00 |
| 1992 | 25,00 | 22,00 |
| 1993 | 25,00 | 22,00 |
| 1994 | 25,00 | 22,00 |
| 1995 | 25,00 | 22,00 |
| 1996 | 25,00 | 22,00 |
| 1997 | 25,00 | 22,00 |
| 1998 | 25,00 | 22,00 |
| 1999 | 25,00 | 22,00 |
| 2000 | 25,00 | 22,00 |
| 2001 | 25,00 | 22,00 |
| 2002 | 25,00 | 22,00 |
| 2003 | 25,00 | 22,00 |
| 2004 | 25,00 | 22,00 |
| 2005 | 25,00 | 22,00 |

TABLEAU 5.8 – Âge de début de carrière

# Annexe D : Table de mortalité INSEE 2014 - 2016

| Âge x | Homme | | Femme | |
|---|---|---|---|---|
| | Survivants $l_x$ à l'âge x | Espérance de vie $E_x$ à l'âge x | Survivants $l_x$ à l'âge x | Espérance de vie $E_x$ à l'âge x |
| 0 | 100 000 | 79,15 | 100 000 | 85,26 |
| 1 | 99 593 | 78,47 | 99 659 | 84,55 |
| 2 | 99 563 | 77,50 | 99 631 | 83,57 |
| 3 | 99 544 | 76,51 | 99 615 | 82,58 |
| 4 | 99 529 | 75,52 | 99 604 | 81,59 |
| 5 | 99 517 | 74,53 | 99 595 | 80,60 |
| 6 | 99 506 | 73,54 | 99 587 | 79,61 |
| 7 | 99 496 | 72,55 | 99 580 | 78,61 |
| 8 | 99 487 | 71,56 | 99 573 | 77,62 |
| 9 | 99 478 | 70,56 | 99 566 | 76,62 |
| 10 | 99 470 | 69,57 | 99 559 | 75,63 |
| 11 | 99 463 | 68,57 | 99 552 | 74,63 |
| 12 | 99 455 | 67,58 | 99 545 | 73,64 |
| 13 | 99 446 | 66,58 | 99 538 | 72,64 |
| 14 | 99 434 | 65,59 | 99 531 | 71,65 |
| 15 | 99 419 | 64,60 | 99 521 | 70,66 |
| 16 | 99 400 | 63,61 | 99 508 | 69,67 |
| 17 | 99 376 | 62,63 | 99 494 | 68,68 |
| 18 | 99 343 | 61,65 | 99 481 | 67,68 |
| 19 | 99 300 | 60,68 | 99 462 | 66,70 |
| 20 | 99 249 | 59,71 | 99 443 | 65,71 |
| 21 | 99 195 | 58,74 | 99 421 | 64,72 |
| 22 | 99 135 | 57,77 | 99 400 | 63,74 |
| 23 | 99 077 | 56,81 | 99 380 | 62,75 |
| 24 | 99 014 | 55,84 | 99 357 | 61,76 |
| 25 | 98 950 | 54,88 | 99 336 | 60,78 |
| 26 | 98 883 | 53,92 | 99 312 | 59,79 |
| 27 | 98 811 | 52,96 | 99 288 | 58,81 |
| 28 | 98 739 | 51,99 | 99 263 | 57,82 |
| 29 | 98 664 | 51,03 | 99 235 | 56,84 |
| 30 | 98 583 | 50,07 | 99 203 | 55,86 |
| 31 | 98 503 | 49,12 | 99 175 | 54,87 |
| 32 | 98 420 | 48,16 | 99 140 | 53,89 |
| 33 | 98 333 | 47,20 | 99 105 | 52,91 |
| 34 | 98 244 | 46,24 | 99 067 | 51,93 |
| 35 | 98 143 | 45,29 | 99 028 | 50,95 |
| 36 | 98 044 | 44,33 | 98 985 | 49,97 |
| 37 | 97 939 | 43,38 | 98 935 | 49,00 |
| 38 | 97 821 | 42,43 | 98 878 | 48,02 |
| 39 | 97 697 | 41,48 | 98 817 | 47,05 |
| 40 | 97 557 | 40,54 | 98 750 | 46,09 |
| 41 | 97 413 | 39,60 | 98 677 | 45,12 |
| 42 | 97 258 | 38,67 | 98 596 | 44,16 |
| 43 | 97 085 | 37,73 | 98 510 | 43,19 |
| 44 | 96 889 | 36,81 | 98 405 | 42,24 |
| 45 | 96 676 | 35,89 | 98 291 | 41,29 |
| 46 | 96 438 | 34,98 | 98 163 | 40,34 |
| 47 | 96 178 | 34,07 | 98 026 | 39,40 |
| 48 | 95 886 | 33,17 | 97 871 | 38,46 |
| 49 | 95 574 | 32,28 | 97 699 | 37,53 |
| 50 | 95 229 | 31,39 | 97 517 | 36,59 |
| 51 | 94 860 | 30,51 | 97 315 | 35,67 |
| 52 | 94 458 | 29,64 | 97 098 | 34,75 |

| Âge x | Homme | | Femme | |
|---|---|---|---|---|
| | Survivants $l_x$ à l'âge x | Espérance de vie $E_x$ à l'âge x | Survivants $l_x$ à l'âge x | Espérance de vie $E_x$ à l'âge x |
| 53 | 94 014 | 28,78 | 96 868 | 33,83 |
| 54 | 93 524 | 27,93 | 96 618 | 32,92 |
| 55 | 92 978 | 27,09 | 96 343 | 32,01 |
| 56 | 92 381 | 26,26 | 96 048 | 31,11 |
| 57 | 91 718 | 25,45 | 95 730 | 30,21 |
| 58 | 91 004 | 24,64 | 95 385 | 29,31 |
| 59 | 90 237 | 23,85 | 95 024 | 28,42 |
| 60 | 89 420 | 23,06 | 94 628 | 27,54 |
| 61 | 88 540 | 22,28 | 94 214 | 26,66 |
| 62 | 87 598 | 21,52 | 93 776 | 25,78 |
| 63 | 86 596 | 20,76 | 93 309 | 24,91 |
| 64 | 85 552 | 20,01 | 92 811 | 24,04 |
| 65 | 84 467 | 19,26 | 92 282 | 23,17 |
| 66 | 83 335 | 18,51 | 91 718 | 22,31 |
| 67 | 82 146 | 17,77 | 91 131 | 21,45 |
| 68 | 80 915 | 17,04 | 90 512 | 20,60 |
| 69 | 79 605 | 16,31 | 89 843 | 19,75 |
| 70 | 78 214 | 15,59 | 89 118 | 18,90 |
| 71 | 76 742 | 14,88 | 88 326 | 18,07 |
| 72 | 75 201 | 14,18 | 87 465 | 17,24 |
| 73 | 73 571 | 13,48 | 86 517 | 16,42 |
| 74 | 71 853 | 12,79 | 85 513 | 15,61 |
| 75 | 70 022 | 12,11 | 84 413 | 14,81 |
| 76 | 68 075 | 11,44 | 83 205 | 14,02 |
| 77 | 66 018 | 10,78 | 81 854 | 13,24 |
| 78 | 63 807 | 10,14 | 80 378 | 12,47 |
| 79 | 61 400 | 9,52 | 78 733 | 11,72 |
| 80 | 58 822 | 8,91 | 76 916 | 10,99 |
| 81 | 56 062 | 8,33 | 74 862 | 10,28 |
| 82 | 53 073 | 7,77 | 72 538 | 9,59 |
| 83 | 49 921 | 7,23 | 69 981 | 8,92 |
| 84 | 46 568 | 6,71 | 67 098 | 8,28 |
| 85 | 43 004 | 6,22 | 63 896 | 7,67 |
| 86 | 39 285 | 5,77 | 60 359 | 7,09 |
| 87 | 35 456 | 5,34 | 56 461 | 6,55 |
| 88 | 31 509 | 4,94 | 52 246 | 6,04 |
| 89 | 27 584 | 4,57 | 47 797 | 5,55 |
| 90 | 23 744 | 4,23 | 43 098 | 5,10 |
| 91 | 20 082 | 3,91 | 38 211 | 4,69 |
| 92 | 16 657 | 3,61 | 33 296 | 4,31 |
| 93 | 13 465 | 3,35 | 28 490 | 3,95 |
| 94 | 10 591 | 3,13 | 23 817 | 3,63 |
| 95 | 8 120 | 2,93 | 19 385 | 3,35 |
| 96 | 6 083 | 2,74 | 15 385 | 3,09 |
| 97 | 4 392 | 2,60 | 11 841 | 2,86 |
| 98 | 3 055 | 2,52 | 8 866 | 2,66 |
| 99 | 2 080 | 2,46 | 6 418 | 2,48 |
| 100 | 1 383 | 2,45 | 4 532 | 2,31 |
| 101 | 896 | 2,52 | 3 051 | 2,18 |
| 102 | 569 | 2,67 | 1 976 | 2,10 |
| 103 | 356 | 2,97 | 1 241 | 2,04 |
| 104 | 234 | 3,26 | 764 | 2,00 |

Tableau 5.9 – Table de de mortalité INSEE 2014 - 2016

# Annexe E : Corrélations parmi les variables

| | Resignation | Gender | MaritalStatus | Category | ExeNonExe | JobLevel | TypeOfEmployment | RTTRight | Entity | City | Region |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Resignation** | 1,00000 | 0,05035 | 0,11784 | 0,11765 | 0,09403 | 0,09567 | 0,02036 | 0,00779 | 0,03824 | 0,12459 | 0,04161 |
| **Gender** | 0,05035 | 1,00000 | 0,12130 | 0,17540 | 0,06644 | 0,10089 | 0,06089 | 0,00037 | 0,25533 | 0,26985 | 0,16887 |
| **MaritalStatus** | 0,11784 | 0,12130 | 1,00000 | 0,08978 | 0,10269 | 0,07634 | 0,02474 | 0,07301 | 0,06797 | 0,09068 | 0,05993 |
| **Category** | 0,11765 | 0,17540 | 0,08978 | 1,00000 | 1,00000 | 1,00000 | 0,46481 | 0,45236 | 0,42494 | 0,29747 | 0,20051 |
| **ExeNonExe** | 0,09403 | 0,06644 | 0,10269 | 1,00000 | 1,00000 | 1,00000 | 0,11909 | 0,24160 | 0,42787 | 0,49005 | 0,34320 |
| **JobLevel** | 0,09567 | 0,10089 | 0,07634 | 1,00000 | 1,00000 | 1,00000 | 0,05955 | 0,25533 | 0,25031 | 0,35322 | 0,20848 |
| **TypeOfEmployment** | 0,02036 | 0,06089 | 0,02474 | 0,46481 | 0,11909 | 0,05955 | 1,00000 | 0,13883 | 0,13698 | 0,09760 | 0,03885 |
| **RTTRight** | 0,00779 | 0,00037 | 0,07301 | 0,45236 | 0,24160 | 0,25533 | 0,13883 | 1,00000 | 0,43919 | 0,61049 | 0,46453 |
| **Entity** | 0,03824 | 0,25533 | 0,06797 | 0,42494 | 0,42787 | 0,25031 | 0,13698 | 0,43919 | 1,00000 | 0,72853 | 0,40592 |
| **City** | 0,12459 | 0,26985 | 0,09068 | 0,29747 | 0,49005 | 0,35322 | 0,09760 | 0,61049 | 0,72853 | 1,00000 | 1,00000 |
| **Region** | 0,04161 | 0,16887 | 0,05993 | 0,20051 | 0,34320 | 0,20848 | 0,03885 | 0,46453 | 0,40592 | 1,00000 | 1,00000 |

TABLEAU 5.10 – Corrélation V-Cramér entre les variables explicatives catégorielles et la variable de réponse

| | Resignation | Age | YearsAtGroup | TotalWorkingYears | TheoreticalSalary | RestatedSalary | MonthlySalary | PercentRVI | BonusRVI | ExceptionalBonus | NumberOfSalaryMonths | PercentRemuneration |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Resignation** | 1,00000 | -0,16849 | -0,15183 | -0,18141 | -0,01458 | -0,03102 | 0,00101 | 0,01316 | -0,00824 | -0,00456 | -0,09765 | -0,01740 |
| **Age** | -0,16849 | 1,00000 | 0,72891 | 0,98398 | 0,29895 | 0,20413 | 0,26931 | 0,16185 | 0,11707 | 0,05183 | -0,13702 | -0,00246 |
| **YearsAtGroup** | -0,15183 | 0,72891 | 1,00000 | 0,75809 | 0,12102 | 0,08283 | 0,07751 | -0,01463 | 0,00916 | 0,02076 | -0,08642 | -0,03282 |
| **TotalWorkingYears** | -0,18141 | 0,98398 | 0,75809 | 1,00000 | 0,20981 | 0,14510 | 0,16913 | 0,04646 | 0,07110 | 0,04505 | -0,15313 | -0,02122 |
| **TheoreticalSalary** | -0,01458 | 0,29895 | 0,12102 | 0,20981 | 1,00000 | 0,87366 | 0,99016 | 0,85838 | 0,78173 | 0,38953 | 0,05337 | 0,09004 |
| **RestatedSalary** | -0,03102 | 0,20413 | 0,08283 | 0,14510 | 0,87366 | 1,00000 | 0,85848 | 0,70832 | 0,89870 | 0,71224 | 0,15563 | 0,10376 |
| **MonthlySalary** | 0,00101 | 0,26931 | 0,07751 | 0,16913 | 0,99016 | 0,85848 | 1,00000 | 0,88702 | 0,77076 | 0,37318 | 0,05078 | 0,10084 |
| **PercentRVI** | 0,01316 | 0,16185 | -0,01463 | 0,04646 | 0,85838 | 0,70832 | 0,88702 | 1,00000 | 0,64687 | 0,25272 | 0,07897 | 0,11971 |
| **BonusRVI** | -0,00824 | 0,11707 | 0,00916 | 0,07110 | 0,78173 | 0,89870 | 0,77076 | 0,64687 | 1,00000 | 0,52762 | 0,06391 | 0,05543 |
| **ExceptionalBonus** | -0,00456 | 0,05183 | 0,02076 | 0,04505 | 0,38953 | 0,71224 | 0,37318 | 0,25272 | 0,52762 | 1,00000 | -0,02290 | 0,01292 |
| **NumberOfSalaryMonths** | -0,09765 | -0,13702 | -0,08642 | -0,15313 | 0,05337 | 0,15563 | 0,05078 | 0,07897 | 0,06391 | -0,02290 | 1,00000 | 0,34528 |
| **PercentRemuneration** | -0,01740 | -0,00246 | -0,03282 | -0,02122 | 0,09004 | 0,10376 | 0,10084 | 0,11971 | 0,05543 | 0,01292 | 0,34528 | 1,00000 |

TABLEAU 5.11 – Corrélation Pearson entre les variables explicatives numériques et la variable de réponse

# Annexe F : Statistiques et traitements des variables explicatives catégorielles

***Gender :***

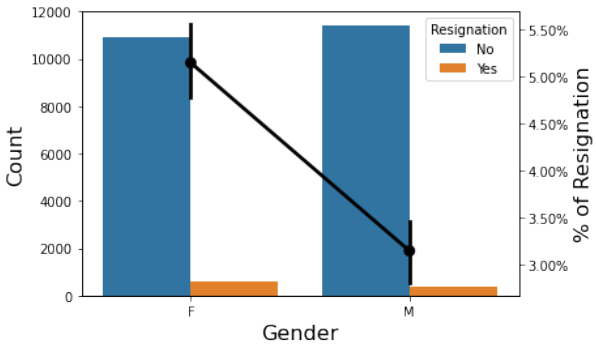
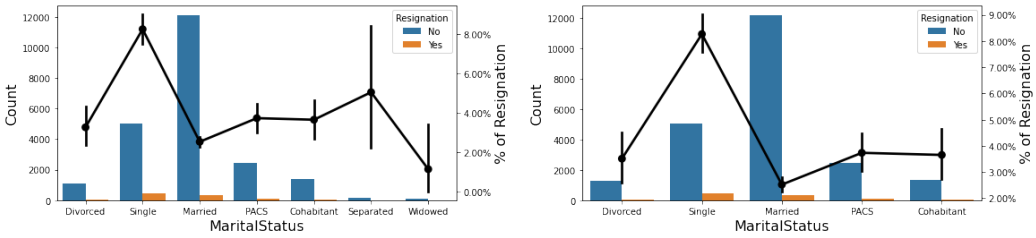



FIGURE 5.13 – Les statistiques et les taux de turnover par Sexe

***MaritalStatus :***




(A) Avant le traitement (B) Après le traitement

FIGURE 5.14 – Les statistiques et les taux de turnover par Statut de mariage avant et après le traitement

***Category :***

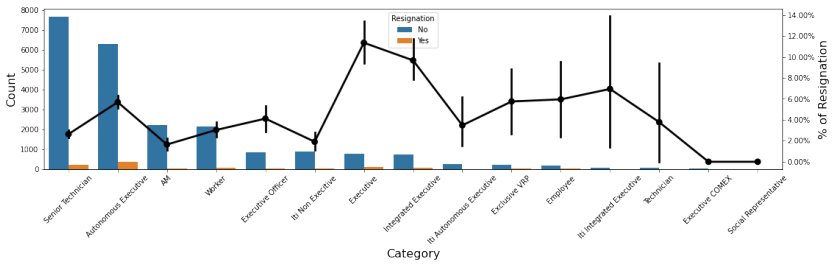



FIGURE 5.15 – Les statistiques et les taux de turnover par Catégorie avant le traitement

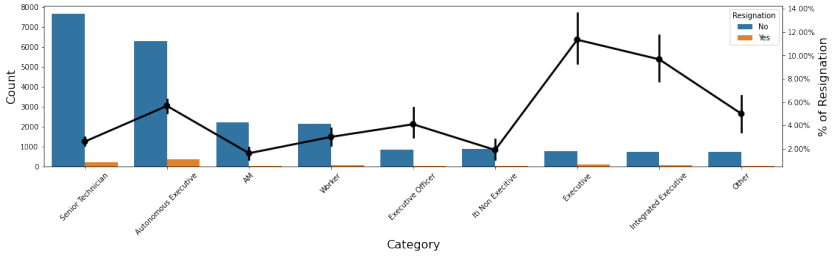



FIGURE 5.16 – Les statistiques et les taux de turnover par Catégorie après le traitement

***ExeNonExe :***

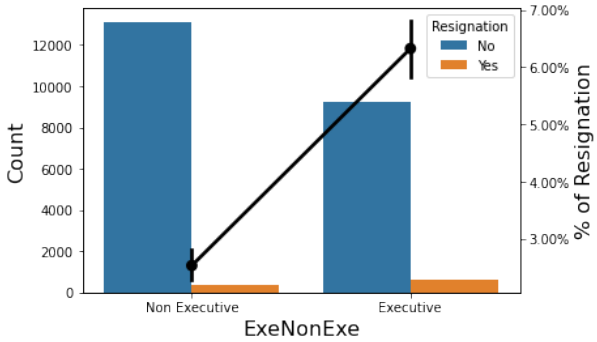



FIGURE 5.17 – Les statistiques et les taux de turnover des salariés Cadres et Non cadres

***Joblevel :***

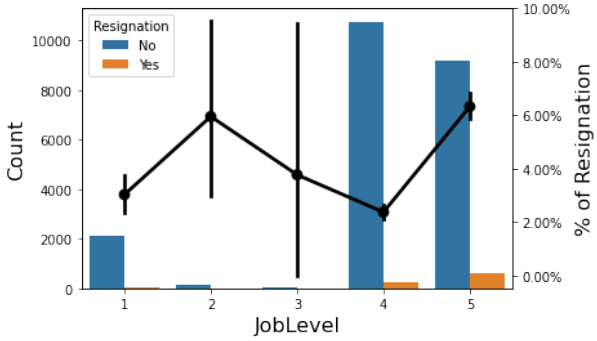



FIGURE 5.18 – Les statistiques et les taux de turnover par Niveau d'emploi

***RTT :***

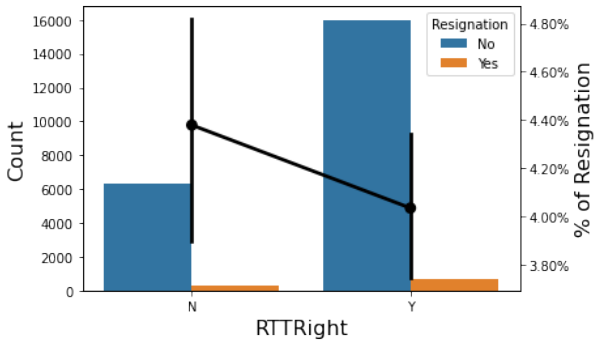



FIGURE 5.19 – Les statistiques et les taux de turnover par Droits de RTT

***City, Region :***

Les traitements de regroupement et de fusion des modalités des variables la *City* et la *Region* sont comme suit :

- *HighZone* : Bretagne Ploermel, Grand Est Strasbourg, Ile-de-France Croissy Beaubourg, Occitanie Toulouse, Ile-de-France St Germain en Laye ;
- *LowZone* : Normandie Lisieux, Occitanie Aramon, Auvergne-Rhone-Alpes Neuville, Occitanie Toulouse, Ile-de-France St Germain en Laye ;
- *ZeroZone* : Nouvelle-Aquitaine Saint Loubes, Nouvelle-Aquitaine Mourenx ;
- *IdF_1* : Ile-de-France Maisons-Alfort, Ile-de-France Paris, Ile-de-France Massy ;
- *IdF_2* : Ile-de-France Marly la ville, Ile-de-France Vitry, Ile-de-France Croix de Berny, Hauts-de-France Compiegne ;
- *IdF_3* : Ile-de-France Chilly-Mazarin, Ile-de-France Gentilly ;
- *CVL_ARA_Nor_Nou* : Centre-Val de Loire Amilly, Centre-Val de Loire Tours, Auvergne-Rhone-Alpes Vertolaye, Normandie Carteret, Nouvelle-Aquitaine Ambares ;
- *Normandie* : Normandie Le Trait, Normandie Val de Reuil, Normandie Elbeuf ;
- *NouvAquitaine* : Nouvelle-Aquitaine Floirac, Occitanie Montpellier ;
- *Lyon* : Auvergne-Rhone-Alpes Lyon ;
- *Marcy l'Etoile* : Auvergne-Rhone-Alpes Marcy l'Etoile ;
- *PACA___Sisteron* : PACA___Sisteron.

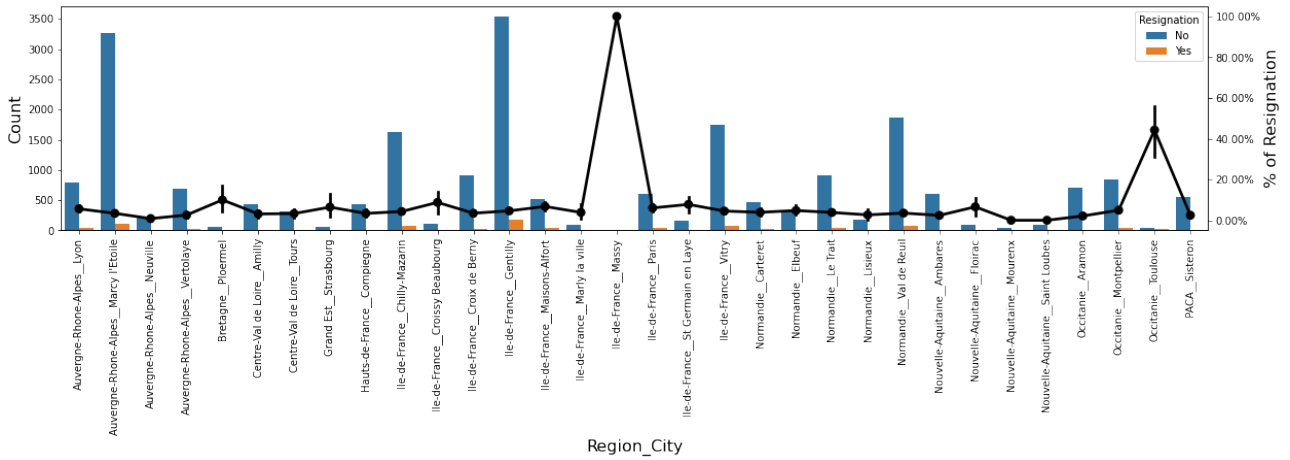



FIGURE 5.20 – Les statistiques et les taux de turnover par Région et Ville avant le traitement

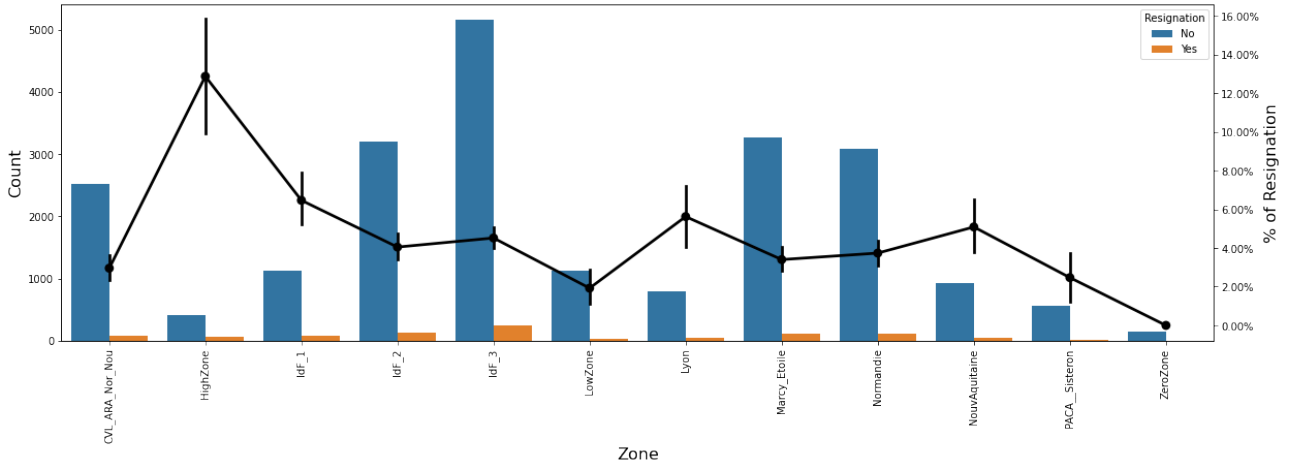



FIGURE 5.21 – Les statistiques et les taux de turnover par Région et Ville après le traitement

# Annexe G : Nouvelle table de turnover prospective

| Année \ Age | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0,020 | 0,024 | 0,024 | 0,030 | 0,037 | 0,043 | 0,044 | 0,050 | 0,049 | 0,054 | 0,052 | 0,056 | 0,057 | 0,058 | 0,061 | 0,059 | 0,052 | 0,052 | 0,051 | 0,048 | 0,046 | 0,039 | 0,036 | 0,030 | 0,026 | 0,024 | 0,023 | 0,021 | 0,018 | 0,018 | 0,016 | 0,014 | 0,012 | 0,011 | 0,011 | 0,009 | 0,009 | 0,009 | 0,009 | 0,008 | 0,007 | 0,007 | 0,007 | 0,006 | 0,006 | 0,005 | 0,005 | 0,004 |
| **1** | 0,023 | 0,030 | 0,031 | 0,036 | 0,044 | 0,053 | 0,054 | 0,058 | 0,056 | 0,059 | 0,058 | 0,060 | 0,060 | 0,058 | 0,061 | 0,055 | 0,049 | 0,049 | 0,046 | 0,045 | 0,043 | 0,035 | 0,031 | 0,026 | 0,022 | 0,022 | 0,020 | 0,019 | 0,016 | 0,016 | 0,015 | 0,012 | 0,011 | 0,010 | 0,010 | 0,009 | 0,009 | 0,009 | 0,009 | 0,008 | 0,006 | 0,006 | | | | | | |
| **2** | 0,029 | 0,036 | 0,037 | 0,044 | 0,053 | 0,064 | 0,063 | 0,066 | 0,063 | 0,067 | 0,063 | 0,064 | 0,060 | 0,058 | 0,059 | 0,052 | 0,045 | 0,044 | 0,042 | 0,043 | 0,038 | 0,032 | 0,029 | 0,022 | 0,020 | 0,019 | 0,018 | 0,017 | 0,015 | 0,015 | 0,013 | 0,011 | 0,010 | 0,009 | 0,009 | 0,009 | 0,009 | 0,008 | 0,009 | 0,007 | 0,005 | | | | | | |
| **3** | 0,034 | 0,042 | 0,044 | 0,053 | 0,065 | 0,072 | 0,070 | 0,070 | 0,068 | 0,071 | 0,065 | 0,065 | 0,059 | 0,055 | 0,057 | 0,048 | 0,041 | 0,041 | 0,040 | 0,039 | 0,033 | 0,029 | 0,024 | 0,020 | 0,017 | 0,018 | 0,016 | 0,016 | 0,014 | 0,012 | 0,011 | 0,010 | 0,009 | 0,009 | 0,009 | 0,008 | 0,008 | 0,008 | 0,008 | 0,006 | | | | | | | | |
| **4** | 0,042 | 0,047 | 0,052 | 0,062 | 0,071 | 0,080 | 0,073 | 0,071 | 0,069 | 0,069 | 0,063 | 0,063 | 0,056 | 0,052 | 0,052 | 0,044 | 0,037 | 0,039 | 0,035 | 0,034 | 0,030 | 0,025 | 0,022 | 0,017 | 0,016 | 0,016 | 0,015 | 0,015 | 0,012 | 0,011 | 0,010 | 0,009 | 0,009 | 0,009 | 0,007 | 0,008 | 0,008 | 0,008 | 0,006 | | | | | | | | | |
| **5** | 0,053 | 0,058 | 0,056 | 0,068 | 0,074 | 0,084 | 0,076 | 0,071 | 0,067 | 0,068 | 0,062 | 0,060 | 0,053 | 0,048 | 0,048 | 0,041 | 0,036 | 0,034 | 0,031 | 0,032 | 0,026 | 0,023 | 0,019 | 0,016 | 0,015 | 0,015 | 0,015 | 0,013 | 0,011 | 0,010 | 0,009 | 0,008 | 0,009 | 0,008 | 0,006 | 0,008 | 0,007 | 0,006 | | | | | | | | | | |
| **6** | 0,071 | 0,066 | 0,055 | 0,072 | 0,073 | 0,083 | 0,076 | 0,071 | 0,066 | 0,067 | 0,057 | 0,056 | 0,049 | 0,043 | 0,045 | 0,040 | 0,032 | 0,029 | 0,028 | 0,027 | 0,023 | 0,020 | 0,018 | 0,015 | 0,013 | 0,014 | 0,013 | 0,012 | 0,010 | 0,009 | 0,009 | 0,008 | 0,009 | 0,008 | 0,006 | 0,007 | 0,006 | | | | | | | | | | | |
| **7** | 0,076 | 0,067 | 0,058 | 0,073 | 0,076 | 0,083 | 0,072 | 0,069 | 0,065 | 0,063 | 0,053 | 0,051 | 0,045 | 0,040 | 0,042 | 0,036 | 0,027 | 0,025 | 0,024 | 0,025 | 0,021 | 0,019 | 0,017 | 0,014 | 0,013 | 0,012 | 0,011 | 0,011 | 0,009 | 0,008 | 0,008 | 0,008 | 0,009 | 0,006 | 0,005 | 0,007 | 0,005 | | | | | | | | | | | |
| **8** | 0,075 | 0,073 | 0,065 | 0,078 | 0,075 | 0,082 | 0,070 | 0,067 | 0,060 | 0,061 | 0,049 | 0,046 | 0,042 | 0,037 | 0,038 | 0,031 | 0,024 | 0,021 | 0,022 | 0,022 | 0,019 | 0,018 | 0,015 | 0,014 | 0,011 | 0,011 | 0,010 | 0,010 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,006 | 0,005 | 0,006 | | | | | | | | | | | | |
| **9** | 0,079 | 0,078 | 0,068 | 0,079 | 0,074 | 0,082 | 0,067 | 0,063 | 0,056 | 0,055 | 0,045 | 0,043 | 0,040 | 0,033 | 0,034 | 0,029 | 0,021 | 0,020 | 0,020 | 0,021 | 0,018 | 0,016 | 0,015 | 0,012 | 0,010 | 0,010 | 0,009 | 0,009 | 0,008 | 0,008 | 0,008 | 0,007 | 0,008 | 0,004 | 0,004 | | | | | | | | | | | | | |
| **10** | 0,094 | 0,075 | 0,064 | 0,079 | 0,068 | 0,077 | 0,059 | 0,058 | 0,048 | 0,050 | 0,041 | 0,040 | 0,035 | 0,028 | 0,030 | 0,026 | 0,019 | 0,018 | 0,018 | 0,019 | 0,016 | 0,016 | 0,013 | 0,010 | 0,010 | 0,009 | 0,008 | 0,008 | 0,008 | 0,008 | 0,007 | 0,007 | 0,007 | 0,004 | | | | | | | | | | | | | | |
| **11** | 0,105 | 0,079 | 0,055 | 0,071 | 0,060 | 0,068 | 0,054 | 0,052 | 0,042 | 0,047 | 0,039 | 0,036 | 0,030 | 0,025 | 0,026 | 0,023 | 0,017 | 0,017 | 0,017 | 0,017 | 0,015 | 0,014 | 0,011 | 0,010 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,007 | 0,007 | 0,007 | 0,007 | | | | | | | | | | | | | | | |
| **12** | 0,088 | 0,078 | 0,049 | 0,066 | 0,057 | 0,064 | 0,048 | 0,048 | 0,038 | 0,045 | 0,035 | 0,030 | 0,026 | 0,021 | 0,023 | 0,020 | 0,016 | 0,015 | 0,016 | 0,016 | 0,013 | 0,012 | 0,010 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,006 | 0,006 | 0,005 | 0,006 | | | | | | | | | | | | | | | |
| **13** | 0,090 | 0,079 | 0,046 | 0,065 | 0,062 | 0,063 | 0,046 | 0,048 | 0,037 | 0,043 | 0,030 | 0,029 | 0,023 | 0,019 | 0,021 | 0,020 | 0,015 | 0,014 | 0,015 | 0,014 | 0,011 | 0,011 | 0,009 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,005 | 0,006 | 0,005 | | | | | | | | | | | | | | | | |
| **14** | 0,113 | 0,094 | 0,044 | 0,069 | 0,060 | 0,065 | 0,047 | 0,049 | 0,035 | 0,040 | 0,029 | 0,025 | 0,021 | 0,018 | 0,020 | 0,019 | 0,015 | 0,014 | 0,013 | 0,012 | 0,011 | 0,010 | 0,008 | 0,007 | 0,008 | 0,008 | 0,008 | 0,007 | 0,007 | 0,004 | 0,006 | | | | | | | | | | | | | | | | | |
| **15** | 0,117 | 0,083 | 0,041 | 0,065 | 0,058 | 0,064 | 0,044 | 0,044 | 0,031 | 0,040 | 0,025 | 0,023 | 0,018 | 0,016 | 0,019 | 0,017 | 0,015 | 0,012 | 0,012 | 0,011 | 0,009 | 0,008 | 0,008 | 0,007 | 0,008 | 0,008 | 0,008 | 0,007 | 0,006 | 0,004 | | | | | | | | | | | | | | | | | | |
| **16** | 0,130 | 0,092 | 0,038 | 0,066 | 0,060 | 0,064 | 0,040 | 0,040 | 0,030 | 0,034 | 0,023 | 0,021 | 0,018 | 0,016 | 0,018 | 0,016 | 0,013 | 0,010 | 0,012 | 0,009 | 0,008 | 0,008 | 0,008 | 0,006 | 0,008 | 0,006 | 0,007 | 0,006 | 0,005 | | | | | | | | | | | | | | | | | | | |
| **17** | 0,102 | 0,089 | 0,036 | 0,064 | 0,057 | 0,057 | 0,034 | 0,035 | 0,025 | 0,030 | 0,020 | 0,020 | 0,017 | 0,015 | 0,017 | 0,014 | 0,011 | 0,010 | 0,010 | 0,008 | 0,008 | 0,008 | 0,008 | 0,006 | 0,007 | 0,006 | 0,006 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **18** | 0,105 | 0,094 | 0,035 | 0,059 | 0,048 | 0,049 | 0,031 | 0,029 | 0,020 | 0,026 | 0,019 | 0,018 | 0,015 | 0,015 | 0,014 | 0,011 | 0,011 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,006 | 0,007 | 0,006 | 0,004 | | | | | | | | | | | | | | | | | | | | |
| **19** | 0,125 | 0,087 | 0,032 | 0,049 | 0,040 | 0,045 | 0,026 | 0,025 | 0,018 | 0,025 | 0,018 | 0,017 | 0,015 | 0,012 | 0,011 | 0,010 | 0,009 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,008 | 0,005 | 0,006 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **20** | 0,103 | 0,077 | 0,022 | 0,038 | 0,034 | 0,040 | 0,022 | 0,023 | 0,017 | 0,022 | 0,016 | 0,017 | 0,012 | 0,010 | 0,011 | 0,009 | 0,008 | 0,007 | 0,008 | 0,008 | 0,008 | 0,007 | 0,007 | 0,005 | 0,006 | | | | | | | | | | | | | | | | | | | | |
| **21** | 0,090 | 0,060 | 0,019 | 0,033 | 0,030 | 0,034 | 0,020 | 0,022 | 0,016 | 0,021 | 0,016 | 0,014 | 0,010 | 0,010 | 0,009 | 0,008 | 0,007 | 0,007 | 0,008 | 0,008 | 0,006 | 0,006 | 0,006 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **22** | 0,062 | 0,047 | 0,016 | 0,028 | 0,027 | 0,029 | 0,018 | 0,020 | 0,014 | 0,018 | 0,013 | 0,011 | 0,009 | 0,008 | 0,008 | 0,007 | 0,007 | 0,006 | 0,008 | 0,007 | 0,005 | 0,006 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **23** | 0,055 | 0,036 | 0,013 | 0,028 | 0,027 | 0,028 | 0,017 | 0,019 | 0,014 | 0,016 | 0,010 | 0,010 | 0,008 | 0,007 | 0,008 | 0,007 | 0,007 | 0,006 | 0,008 | 0,007 | 0,005 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **24** | 0,050 | 0,031 | 0,013 | 0,028 | 0,025 | 0,027 | 0,015 | 0,018 | 0,013 | 0,013 | 0,010 | 0,009 | 0,007 | 0,007 | 0,008 | 0,008 | 0,006 | 0,005 | 0,007 | 0,005 | 0,004 | | | | | | | | | | | | | | | | | | | | |
| **25** | 0,051 | 0,031 | 0,013 | 0,026 | 0,022 | 0,024 | 0,015 | 0,015 | 0,011 | 0,012 | 0,009 | 0,007 | 0,007 | 0,007 | 0,008 | 0,008 | 0,006 | 0,005 | 0,006 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **26** | 0,044 | 0,027 | 0,014 | 0,022 | 0,019 | 0,021 | 0,013 | 0,011 | 0,010 | 0,009 | 0,007 | 0,007 | 0,007 | 0,007 | 0,008 | 0,008 | 0,005 | 0,005 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **27** | 0,035 | 0,026 | 0,013 | 0,018 | 0,017 | 0,016 | 0,010 | 0,010 | 0,008 | 0,008 | 0,007 | 0,007 | 0,007 | 0,007 | 0,008 | 0,007 | 0,004 | 0,005 | | | | | | | | | | | | | | | | | | | | |
| **28** | 0,031 | 0,022 | 0,012 | 0,016 | 0,014 | 0,012 | 0,010 | 0,008 | 0,007 | 0,007 | 0,007 | 0,007 | 0,007 | 0,006 | 0,006 | 0,005 | 0,004 | | | | | | | | | | | | | | | | | | | | | |
| **29** | 0,024 | 0,017 | 0,011 | 0,012 | 0,010 | 0,010 | 0,008 | 0,007 | 0,007 | 0,007 | 0,007 | 0,007 | 0,006 | 0,005 | 0,005 | 0,005 | | | | | | | | | | | | | | | | | | | | | | |
| **30** | 0,020 | 0,015 | 0,008 | 0,009 | 0,008 | 0,009 | 0,007 | 0,007 | 0,007 | 0,007 | 0,007 | 0,005 | 0,006 | 0,005 | 0,005 | | | | | | | | | | | | | | | | | | | | | | | |
| **31** | 0,016 | 0,012 | 0,007 | 0,008 | 0,008 | 0,008 | 0,007 | 0,007 | 0,007 | 0,006 | 0,007 | 0,005 | 0,005 | 0,004 | | | | | | | | | | | | | | | | | | | | | | | | |
| **32** | 0,011 | 0,008 | 0,006 | 0,007 | 0,007 | 0,008 | 0,007 | 0,007 | 0,007 | 0,006 | 0,006 | 0,004 | 0,005 | | | | | | | | | | | | | | | | | | | | | | | | | |
| **33** | 0,010 | 0,008 | 0,007 | 0,007 | 0,007 | 0,008 | 0,007 | 0,007 | 0,007 | 0,005 | 0,005 | 0,004 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **34** | 0,009 | 0,007 | 0,006 | 0,007 | 0,007 | 0,008 | 0,007 | 0,006 | 0,006 | 0,004 | 0,005 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **35** | 0,009 | 0,007 | 0,006 | 0,006 | 0,007 | 0,008 | 0,007 | 0,006 | 0,006 | 0,004 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **36** | 0,009 | 0,007 | 0,006 | 0,006 | 0,006 | 0,008 | 0,006 | 0,005 | 0,005 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **37** | 0,009 | 0,006 | 0,006 | 0,006 | 0,006 | 0,008 | 0,006 | 0,004 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **38** | 0,008 | 0,006 | 0,005 | 0,006 | 0,006 | 0,007 | 0,005 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **39** | 0,008 | 0,006 | 0,005 | 0,005 | 0,005 | 0,006 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **40** | 0,007 | 0,006 | 0,005 | 0,005 | 0,005 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **41** | 0,006 | 0,005 | | 0,005 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **42** | 0,006 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Tableau 5.12 – Taux de turnover de la nouvelle table prospective

# Annexe H : Glossaire des intitulés comptables

| Intitulé (Anglais) | Intitulé (Français) | Définition |
|---|---|---|
| Net Amount Recognized | Provision | Différence entre l'actif de couverture et l'engagement actuariel |
| Defined Benefit Obligation | Engagement actuariel | Valeur actuelle probable des prestations futures sur la période des services rendus |
| Plan Asset | Actif de couverture | Juste valeur des actifs à la date de clôture |
| Service Cost | Coûts des services rendus | Accroissement de l'engagement résultants des services rendus au cours de la période |
| Interest Cost | Coût de l'actualisation | Accroissement de l'engagement résultant du fait que l'on s'est rapproché de la date de règlement des prestations |
| Interest Income | Rendement financier | Rendement des actifs de couverture que l'on compte réaliser sur la période (mesuré avec le taux d'actualisation de la dette) |
| Benefit Payments | Prestations payées | Prestations (diminuant la valeur de l'actif si elles sont prélevées sur le fonds constitué en couverture) |
| Contributions | Contributions employeur | Cotisations versées sur le fonds dédiés à la couverture de l'engagement et augmentant l'actif de couverture |
| Actuarial gains/lossses | Pertes et gains actuariels | Effet des changements d'hypothèses et/ou de la différence entre le résultat hypothétique et la réalité |
| Curtailment (Past Service Cost) | Plan de restructuration (ou coût des services passées) | Intervient lorsqu'une entreprise peut démontrer qu'elle s'est engagée à réduire de façon significative le nombre de personnes bénéficiant d'un régime |
| Settlement | Liquidation | Transaction (autre que le paiement) éliminant toute obligation juridique ou implicite ultérieure pour tout ou partie des prestations prévues par un régime à prestations définies |

Tableau 5.13 – Glossaire des intitulés comptables