Mémoire présenté le :

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

Par : Berthille Pierron

Titre : Modélisation des hypothèses de calcul des passifs sociaux

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

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise : SECOIA

Nom : Hami Sadeck

Signature :

Directeur de mémoire en entreprise :

Nom : Sadeck HAMI

Signature :

Invité :

Nom :

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

Sadeck HAMI

SECOIA
Services Conseil Informatique Actuariat
26 rue Bellecordière - 69002 LYON
Téléphone : 04 78 24 24 73
secoia@secoia.fr
SIRET 389 871 344 00062

Signature du candidat

Mémoire Institut des Actuaires

Modélisation des hypothèses de calcul des passifs sociaux

Auteur :

Berthille Pierron

Tuteur ISFA : Mme Anne Eyraud-Loisel

Tuteur Entreprise : M. Sadeck Hami

Année 2020-2021

Remerciements

Je tiens à remercier l'intégralité de l'équipe passifs sociaux de Secoia pour leur bienveillance à mon égard.

Tout particulièrement, je remercie Monsieur Sadeck Hami pour ses conseils toujours avisés ainsi que ses enseignements.

Il me tient également à cœur de remercier Hana Meskine pour son soutien, ses relectures et, plus important encore, son amitié. Je remercie ensuite l'intégralité de ma famille pour avoir été à mes côtés lors de l'écriture de ce mémoire.

Je remercie, enfin, Madame Anne Eyraud Loisel pour son encouragement ainsi que sa relecture soigneuse.

Résumé

Mots clés : Norme IAS19, passifs sociaux, indemnités de fin de carrière, *Machine Learning*, hypothèses de calcul, modèle linéaire généralisé, R

Les passifs sociaux d'une entreprise, correpondant à la dette de l'employeur envers ses salariés en termes d'avantages, peuvent constituer des montants assez conséquent. Leur évaluation « au plus juste » est alors primordiale.

Ce mémoire a été réalisé dans le cadre de mes missions chez Secoia au sein de l'équipe passifs sociaux. Secoia est un cabinet d'actuariat basé à Lyon et Paris et spécialisé dans le domaine des évaluations d'engagements sociaux. Secoia souhaitait à travers cette étude mettre à profit les données à sa disposition afin de proposer des axes d'amélioration dans le calcul des engagements sociaux. Riche de ces enseignements, l'étude s'est portée sur la mise en place d'un modèle permettant le calibrage des hypothèses de calcul ainsi que le calcul de l'engagement lui-même.

Dans ce mémoire, nous avons donc étudié les hypothèses d'un portefeuille d'une dizaine d'entreprises afin de les calibrer au mieux à l'aide d'une part de modèles non-paramétriques issus du *machine learning* et de modèles paramétriques à l'aide de modèles linéaires généralisés, d'autre part.

Il convenait, dans une première partie, de revenir sur la norme IAS19 régissant la comptabilisation des avantages aux personnels. Ensuite, une étude de sensibilité de ces hypothèses fut menée afin d'identifier les plus impactantes sur l'engagement final de l'entreprise. L'étude s'est donc axée sur la modélisation du *tun-over* et la modélisation de l'évolution des salaires.

Après l'implémentation des modèles sous R, la sélection du modèle le plus performant pour chacune de nos hypothèses s'est faite à l'aide des critères de performances adaptés à chacune des problématiques : la classification binaire pour le *tun-over* et la régression pour la modélisation des salaires.

Enfin, le calcul de l'engagement à l'aide des différents modèles permet de mettre en évidence la pertinence de la mise en place d'une telle approche dans le calcul de l'engagement.

Abstract

Key words : IAS19, social liabilities, leaving service indemnities, Machine Learning, Hypothesys, generalized linear model, R

The company's social liabilities, representing what the employer owes its employees in terms of benefits, can constitute sizeable amounts for the company. Their evaluation as accurately as possible is therefore essential.

This dissertation was produced as part of my assignments at Secoia within the social liabilities team. Secoia, an actuarial firm based in Lyon and Paris, is specialized in the evaluation of social liabilities. Secoia, throught this paper, wanted to profit its data in order to propose areas for improvement in the evaluation of liabilities. This study focused on the implementation of a model allowing the calibration of the assumptions for the evaluation of the liabilities as well as the calculation of the liabilities itself.

In this dissertation, we have therefore studied the hypotheses of a portfolio of ten companies in order to calibrate them as best as possible using non-parametric models through *machine learning* on the one hand and parametric models using generalized linear models on the other hand.

In the first part, it was necessary to come back to the IAS19 standard governing the recognition of employee benefits. Then, a sensitivity study of these hypotheses was carried out in order to identify the most impacting on the company's final liability. The study therefore focused on the modeling of the *tun-over* and the modeling of the evolution of wages.

After implementing the models under R, the selection of the best performing model for each of our hypotheses was made using performance criteria adapted to each of the problematics : binary classification for the turn-over and regression for the wages modelisation.

Finally, the calculation of the liability using the different models makes it possible to highlight the relevance of the implementation of such an approach in the calculation of social liabilities.

Table des matières

Résumé 1

Abstract 2

Introduction 7

I La norme IAS 19 et notre problématique 9

1 Les grands principes de la norme IAS 19 10

1.1 Les normes IFRS . . . 10

1.1.1 Introduction . . . 10

1.1.2 Buts fondamentaux . . . 10

1.2 Présentation de la norme IAS 19 . . . 11

1.2.1 Introduction . . . 11

1.2.2 Objectifs, points clés et enjeux . . . 11

1.2.3 Définition et comptabilisation des passifs sociaux . . . 12

1.3 Evaluation actuarielle . . . 14

1.3.1 Méthodes d'évaluation . . . 16

1.3.2 Hypothèses actuarielles . . . 19

1.4 Comptabilisation . . . 22

1.4.1 La provision . . . 22

1.4.2 La charge . . . 22

1.4.3 Reconnaissance des écarts actuariels . . . 23

1.5 Les évènements spéciaux . . . 25

1.5.1 Modification de régime . . . 25

1.5.2 Réduction de régime . . . 25

1.5.3 Liquidation de régime . . . 26

1.5.4 Transferts . . . 26
1.6 Conclusion du chapitre . . . 26

2 Problématique **28**
2.1 L'enjeu de l'étude . . . 28
2.2 Contexte législatif des indemnités de fin de carrière . . . 29
2.3 RGPD . . . 30
2.4 Constitution de la base de données . . . 30
2.5 Données retenues et statistiques descriptives . . . 31
2.6 Sensibilité de l'engagement aux différentes hypothèses . . . 34
2.6.1 Sensibilité au taux d'actualisation . . . 34
2.6.2 Sensibilité au taux de revalorisation des salaires . . . 35
2.6.3 Sensibilité au taux de mortalité . . . 36
2.6.4 Sensibilité au taux de turn-over . . . 36
2.6.5 Sensibilité au taux de charges sociales . . . 37
2.6.6 Sensibilité à l'âge de départ en retraite . . . 37
2.7 Conclusion du chapitre . . . 38

II Modélisations et résultats **39**

3 La modélisation du turnover **40**
3.1 Les motifs de sortie à intégrer dans notre modélisation . . . 40
3.2 Traitement des données . . . 41
3.3 Analyse de données . . . 42
3.3.1 Quelques statistiques descriptives . . . 42
3.3.2 Analyse en composantes principales . . . 43
3.4 Modélisation . . . 45
3.4.1 Les Modèles linéaires généralisés . . . 46
3.4.2 Le Machine learning . . . 49
3.4.3 L'avant-propos à la construction des modèles . . . 53
3.5 La mise en place de la régression logistique . . . 55
3.5.1 Création du modèle et sélection des variables . . . 56
3.5.2 Mesure de la qualité d'ajustement du modèle . . . 59
3.5.3 Mesures de la qualité prédictive du modèle . . . 61
3.5.4 Les performances de la régression logistique . . . 62

3.6 Implémentation de l'agorithme CART . . . 62
3.6.1 Construction de l'arbre et élaguage . . . 62
3.6.2 Les performances de l'arbre CART . . . 64
3.7 Implémentation du random forest . . . 64
3.7.1 Construction du modèle . . . 64
3.7.2 Les performances du random forest . . . 66
3.8 Implémentation du Gradient Boosting . . . 67
3.8.1 Construction du modèle . . . 67
3.8.2 Les performances du gradient boosting . . . 68
3.9 L'importance des variables explicatives . . . 68
3.10 Le choix du modèle final . . . 69
3.11 Conclusion du chapitre . . . 70

4 La modélisation de l'évolution des salaires 72
4.1 Traitement des données . . . 72
4.2 Statistiques descriptives et analyse en composantes principales . . . 73
4.2.1 Statistiques descriptives . . . 73
4.2.2 Analyse en composantes principales . . . 73
4.3 L'inflation est-elle à prendre en compte dans l'évolution des salaires ? . . . 76
4.3.1 Corrélation de Pearson . . . 77
4.3.2 Corrélation de Spearman . . . 78
4.3.3 Présence d'ex-aequo . . . 78
4.3.4 Applications . . . 78
4.4 Modélisation du taux d'augmentation des salaires . . . 79
4.4.1 Modèle linéaire généralisé loi Gamma . . . 80
4.4.2 Création du modèle et sélection des variables . . . 80
4.4.3 Validation du modèle . . . 82
4.4.4 Performance prédictive du modèle . . . 83
4.4.5 La régression en apprentissage automatique . . . 84
4.5 Importance des variables dans la modélisation de l'évolution des salaires . . . 85
4.6 Le choix du modèle d'évolution des salaires final . . . 86
4.7 Implémentation d'une probabilité d'être promu . . . 86
4.7.1 Le déséquilibre . . . 89
4.7.2 Correction algorithmique . . . 90
4.7.3 Correction au niveau des données . . . 91

4.8 Application de la correction du déséquilibre . . . 92
4.8.1 Régression logistique . . . 92
4.8.2 Arbre de décision . . . 93
4.8.3 Random Forest . . . 94
4.8.4 Gradient Boosting . . . 94
4.9 Importance des variables dans la modélisation des promotions . . . 95
4.10 Le choix final du modèle promotion . . . 96
4.11 Conclusion du chapitre . . . 97

5 Prise en compte dans le calcul de l'engagement 98
5.1 Le modèle « PUC service prorata » . . . 98
5.2 Les projections de la probabilité de turn-over . . . 99
5.3 Les résultats suite à la prise en compte du modèle de turn-over . . . 100
5.4 Les salaires projetés . . . 101
5.5 Les résultats suite à la prise en compte du modèle d'évolution des salaires . . . 102
5.6 Les résultats suite à la prise en compte des promotions . . . 103
5.6.1 Critiques des modèles . . . 104
5.7 Recalcul des passifs sociaux . . . 105
5.7.1 Engagements au 31/12/2019 . . . 106
5.7.2 Engagements au 31/12/2020 . . . 106
5.7.3 Evolution de l'engagement . . . 106

Conclusion 107

Bibliographie 110

Annexes 111

Table des figures 120

Table des figures 123

Liste des tableaux 125

Introduction

Les avantages aux personnels correspondent à « toutes formes de contrepartie donnée par une entreprise au titre des services rendus par son personnel ». Ainsi, l'entreprise reconnait une « dette » vis-à-vis de ses salariés. La reconnaissance de cette dette est régie par la norme IAS19 qui impose à toutes sociétés cotées en Europe la comptabilisation de celle-ci dans leurs comptes.

Ceci engendre le chiffrage de cette dette qui se fait à l'aide d'un certain nombre d'hypothèses. L'objectif de ce mémoire est la modélisation de ces hypothèses.

Il conviendra, dans une première partie, de revenir sur la norme IAS19 afin d'appréhender les tenants et aboutissants du calcul de ces passifs sociaux. Ensuite, une étude de sensibilité de ces hypothèses sera menée afin d'identifier les hypothèses ayant le plus d'impact sur l'engagement de l'entreprise.

Par la suite, les premiers travaux à entreprendre avant la modélisation seront la constitution de la base de données. Enfin, après l'implémentation de nos différents modèles et la sélection des modèles les plus performants, nous analyserons les résultats afin de mettre en évidence la pertinence des modélisations réalisées.

Première partie

La norme IAS 19 et notre problématique

Chapitre 1

Les grands principes de la norme IAS 19

Dans ce chapitre nous allons tout d'abord, nous intéresser aux normes IFRS et leur utilité. Puis, nous allons nous intéresser plus précisément à la norme IAS19, l'objet de ce mémoire, afin de comprendre les tenants et aboutissants de l'évaluation des passifs sociaux, de leur identification jusqu'à leur comptabilisation.

1.1 Les normes IFRS

1.1.1 Introduction

Concernant les normes IFRS (*International Financial Reporting Standards*), celles-ci sont établies par un organisme supranational composé de différents experts indépendants venant des quatre coins du monde réunis au sein de l'*International Accounting Standards Board* (IASB). Créé initialement en 1973, cet organisme est en charge d'élaborer les normes comptables internationales dites IFRS ou encore les *International Accounting Standards* (IAS).

A ce jour, ce jeu de normes regroupe 45 textes actuellement en vigueur et couvrent un champ extrêmement vaste de domaines allant, par exemple, de la comptabilisation des contrats d'assurance aux impôts sur le revenu. Ces normes ont été adoptées dans plus de 144 pays à travers le monde, notamment au sein de l'Union Européenne. Effectivement, en vertu du règlement CE 1606/2002, toutes sociétés cotées sur les marchés réglementés en Europe ont l'obligation, depuis le 1er janvier 2005, de publier leurs comptes consolidés suivant les normes IFRS. En France, les groupes non cotés peuvent aussi, au choix, opter pour les normes IFRS ou conserver les règles de comptabilité Française éditées quant à elles par l'ANC (Autorité des Normes Comptables).

1.1.2 Buts fondamentaux

Les normes produites par l'IASB sont destinées à standardiser et ainsi à harmoniser la présentation des états comptables au niveau mondial. Le but est de donner le plus de clarté possible aux comptes ainsi que plus de transparence en évaluant au mieux la performance financière des entreprises. L'adoption de ces normes au niveau international permet d'accroître,

par exemple, la comparabilité des états financiers de sociétés d'un même secteur mais provenant de pays différents.

1.2 Présentation de la norme IAS 19

1.2.1 Introduction

Nous présentons dans cette partie la norme IAS 19 relative aux « avantages au personnel » faisant l'objet de l'étude. Cette norme n'est pas récente puisque initialement publiée en 1998 puis amendée plusieurs fois depuis. La dernière version, bien qu'ayant fait l'objet de mises à jour minimes depuis, est appelée IAS 19 Révisée et a été adoptée par l'Union Européenne le 5 juin 2012, rendant son application obligatoire pour toutes entreprises cotées à compter du 1er janvier 2013.

A noter que le Code du Commerce Français laisse, quant à lui, le choix aux entreprises non cotées de provisionner ses engagements selon cette norme ou bien d'en indiquer le montant en annexe uniquement.

La norme IAS 19 Révisée doit être appliquée pour la comptabilisation, par l'employeur, de tous les avantages du personnel, sauf ceux auxquels s'applique IFRS 2 relative au « Paiement fondé sur des actions ».

1.2.2 Objectifs, points clés et enjeux

La norme IAS 19 « avantages aux personnels » a pour principal objet d'imposer aux entreprises d'identifier, de valoriser ainsi que de provisionner la charge future probable à laquelle elles seront exposées à raison d'avantages consentis à leur personnel, qu'ils découlent d'obligations juridiques nommées (dispositions légales ou réglementaires, convention collective, accord d'entreprise...) ou d'une obligation implicite innommée dite d'usage.

« L'objectif de la présente norme est de prescrire le traitement comptable et les informations à fournir pour les avantages du personnel. La norme impose à une entité de reconnaître :
(a) un passif lorsqu'un membre du personnel a rendu des services en échange d'avantages aux personnels à payer dans le futur ; et
(b) une charge lorsque l'entité consomme l'avantage économique découlant du service rendu par un membre du personnel en échange d'avantages au personnel. » (paragraphe 1, IAS19 R)

Elle instaure ainsi une méthode d'évaluation commune à toutes les entreprises permettant d'augmenter la comparabilité des sociétés entre elles, un objectif fondamental des normes IFRS.

La norme s'applique à l'ensemble des prestations accordées par une entreprise au titre des services rendus par son personnel y compris à celles versées pendant la durée de vie active. Ceci intègre entre autres les salaires, les congés payés, les médailles du travail, les Comptes Épargne Temps (CET), les indemnités de départ, etc... ce qui représente des sommes considérables d'où l'enjeu d'une évaluation et d'une comptabilisation appropriée.

1.2.3 Définition et comptabilisation des passifs sociaux

La norme IAS 19 décrit pour chacune des catégories d'avantages au personnel les règles d'évaluation et de comptabilisation ainsi que les informations à fournir. La définition des avantages au personnel donnée est la suivante :

« Les avantages du personnel désignent toutes formes de contrepartie versées par une entité en échange de services rendus par des salariés ou en cas de cessation d'emploi. » (paragraphe 8, IAS19 R)

Il existe cinq catégories d'avantages au personnel, ci-après énumérées :

Les avantages à court terme

Les avantages à court terme désignent les avantages (autres que les indemnités de fin de contrat de travail) qui sont intégralement dus dans les 12 mois suivant la fin de la période pendant laquelle les membres du personnel ont rendu les services correspondants. L'entité n'a besoin d'aucune hypothèse actuarielle ou méthode de projection pour mesurer les coûts annuels de ces avantages à court terme. En effet, de par leur échéance courte et leur quasi-certitude de paiement, ces avantages ne génèrent pas de passifs sociaux pour l'entreprise.

Exemples :

— Monétaires : salaires, rémunérations, cotisations de sécurité sociale, congés payés, intéressement, primes, …

— Non monétaires : assistance médicale, logement, voiture, …

Les avantages à long terme

On appelle avantages à long terme, les avantages (autres que les avantages postérieurs à l'emploi et indemnités de fin de contrat de travail) dont le règlement intervient généralement durant la vie active du bénéficiaire et qui ne sont pas dus intégralement dans les 12 mois suivant la fin de la période pendant laquelle les membres du personnel ont rendu les services correspondants. Ces avantages sont comptabilisés sur la base des charges actuarielles. L'entité doit réaliser une évaluation actuarielle, fondée sur des hypothèses d'actualisation, de probabilisation et des méthodes de projection, afin de mesurer son obligation et calculer ses charges annuelles. La nécessité d'actualisation pour ces avantages vient du fait que son paiement intervient plus de douze mois après le service rendu. Quant à la nécessité de probabiliser ce montant, cela vient du fait que certains avantages ne sont acquis qu'au moment de leur paiement.

Exemples : Médailles du travail, primes d'ancienneté, accumulations de congés à long terme, congés sabbatiques, primes payables au-delà de 12 mois.

Les avantages postérieurs à l'emploi

Ils désignent, comme leur nom l'indique, les avantages dont le règlement se fait postérieurement à la cessation de l'emploi. La plupart sont des prestations de retraite et sont caractérisées par le fait d'être versées par la société à ses salariés au moment et pendant leur retraite. Il existe deux catégories d'avantages postérieurs à l'emploi :

Les régimes à cotisations définies : Les régimes à cotisations définies dénomment les régimes pour lesquels une entité verse des cotisations à un tiers (généralement sur un fonds) et n'aura aucune obligation juridique ou implicite, de payer des cotisations supplémentaires. L'obligation de l'employeur est une obligation de moyen : le financement de ces cotisations.

La société comptabilise une charge en contrepartie des montants versés sur le fonds et est libérée de l'obligation, il n'y a donc pas d'engagement à la reconnaitre au bilan sous forme de provision.

Ici, la société ne porte pas de risque puisqu'elle se limite à verser un certain nombre de cotisation généralement défini comme un pourcentage de salaire.

L'entité est donc simplement tenue de comptabiliser les coûts annuels. Elle n'a besoin d'aucune hypothèse actuarielle ou méthode de projection pour mesurer cette obligation. Le coût annuel de l'employeur correspond au montant du versement durant la période comptable.

Exemples :

— France : régimes de retraite « Article 83 », PERCO, …

— International : régime de retraite 401 (k) aux Etats-Unis, fonds de prévoyance en Asie, …

Les régimes à prestations définies : Les régimes à prestations définies se distinguent des régimes à cotisations définies puisqu'ils fournissent, à la retraite, un avantage dont la valeur est définie à l'avance par une formule. L'obligation de l'employeur est, cette fois-ci, une obligation de résultat : montant de prestation connu à l'avance.

La société a une obligation de payer des prestations convenues. Elle reconnait un engagement au bilan correspondant à l'évaluation faite des montants à régler dans le futur à ses salariés au titre des avantages postérieurs à l'emploi.

Ici, seule la société porte le risque car elle s'engage à verser une prestation définie lors de la mise en place du dispositif qui dépend généralement des derniers salaires du salarié.

Ces avantages sont comptabilisés sur la base de charges actuarielles. 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.

Le régime à prestations définies peut prendre deux formes :

— Un régime dit additif : La prestation est définie comme un pourcentage du dernier salaire, indépendamment des retraites des régimes de bases et complémentaires. Ainsi, la prestation dépend uniquement du dernier salaire, du taux de retraite garanti et de l'ancienneté du salarié, d'où son nom puisque ces rentes viennent s'ajouter aux différentes rentes percues par le salarié.

— Un régime dit différentiel : Cette fois-ci, la société garanti un niveau global de retraite tous régimes confondus, la rente versée au titre du dispositif différentiel est calculée sous déduction des autres régimes. On peut dire que l'employeur s'engage à combler la différence entre les retraites de bases et complémentaires et le niveau garanti par le dispositif.

Exemples : Indemnités de départ à la retraite, régimes de retraite « Article 39 », frais de santé des retraités, …

Les indemnités de cessation d'emploi

Ces avantages au personnel sont payables suite à la cessation d'activité. Cette cessation est alors l'élément qui génère l'obligation et peut être décidée par l'entité qui résilie le contrat de travail du membre du personnel avant l'âge normal de départ en retraite ou par le membre du personnel qui prend la décision de partir volontairement en échange de ces indemnités. Les coûts doivent immédiatement être comptabilisés sous forme de charges l'année durant laquelle ils surviennent via une provision de restructuration.

Exemples : Indemnités liées à un licenciement, indemnités de plan de départ volontaire avant l'âge de la retraite, plan de restructuration, …

Les avantages payés en action (IFRS 2)

Ces avantages dont la valeur dépend du prix de l'action de l'entité, sont traités séparément dans la norme IFRS 2.

Exemples : Actions gratuites, stock-options, …

1.3 Evaluation actuarielle

Comme nous venons de le voir, seules certaines catégories d'avantages aux personnels génèrent des passifs sociaux dont la provision doit être calculer via une évaluation actuarielle. Cette évaluation est requise lorsqu'il y a de l'incertitude, c'est cette incertitude qui caractérise principalement la provision :

- Incertitude sur le paiement en lui-même : le décès du salarié libère, de fait, l'entreprise de son engagement. A noter qu'en fonction des accords applicables, le licenciement ou la démission peuvent également libérer l'entreprise de son engagement.
- Incertitude sur la date de paiement : l'âge de départ à la retraite n'est pas figé.
- Incertitude sur le montant à payer : l'indemnité de départ en retraite dépend du niveau de rémunération lors des dernières années travaillées et de l'ancienneté, ainsi que d'accords d'entreprise ou des conventions collectives.

C'est alors le rôle de l'actuaire d'estimer ces paramètres de la manière la plus juste possible afin de livrer une vision fiable quant aux coûts finaux des avantages au personnel.

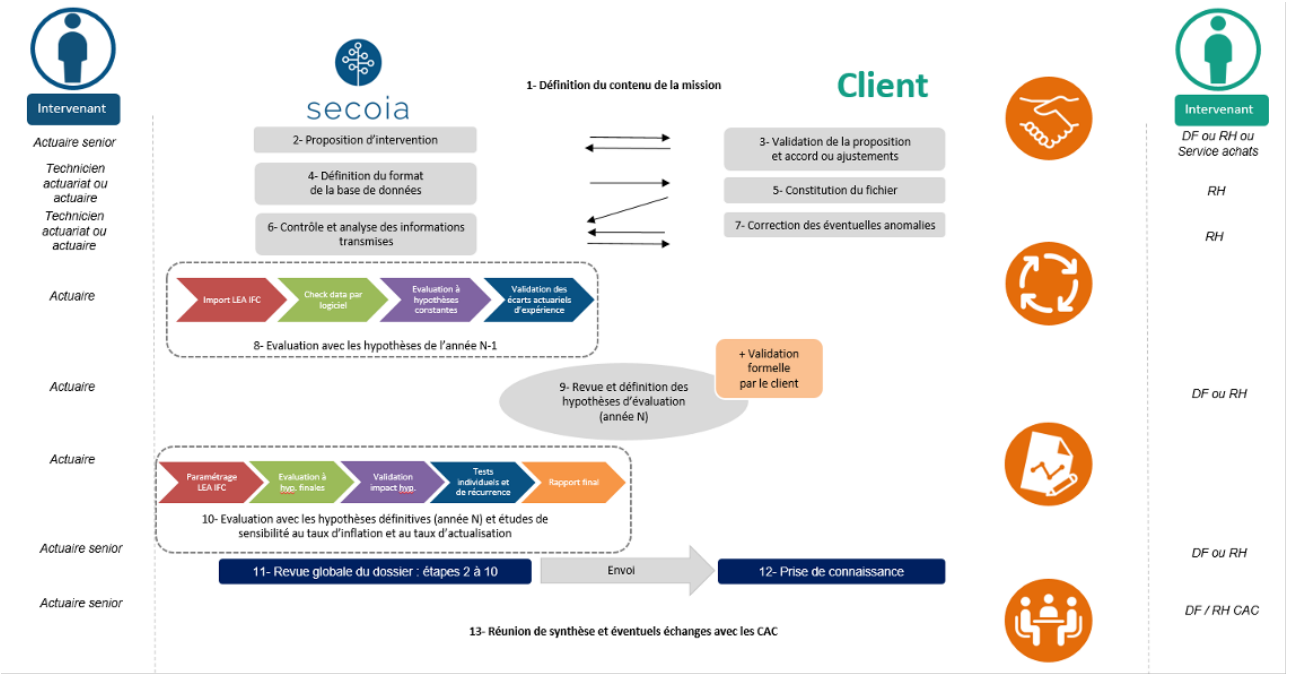


FIGURE 1.1 – Processus d'évaluation des passifs sociaux

Le processus d'évaluation actuarielle des passifs sociaux, schématisé ci-dessus, se compose de plusieurs étapes :

- Etape 1 : la première étape repose sur la constitution d'un cahier des charges délimitant l'intervention de l'actuaire. Lorsque les termes de l'accord entre l'entreprise désirant faire calculer son engagement et le cabinet d'actuaires sont conclus, l'actuaire transmet les éléments dont il a besoin afin de mener à bien la mission.
- Etape 2 : il s'agit de la réception des données transmise par le client. L'actuaire va donc procéder à toute une série de vérifications afin de mettre en évidence des incohérences potentielles d'une part, mais également procèder à une réconciliation des données d'une année à l'autre. En effet, afin de mettre en évidence les différents éléments intervenants dans la diminution ou la hausse de la provision, il est nécessaire d'identifier, par exemple, les entrées et sorties d'effectif et les évolutions des salaires. Généralement s'en suit une série de questions - réponses avec le client pour valider les données utilisées lors de l'évaluation.
- Etape 3 : une fois les données validées, l'actuaire se penche sur les hypothèses de calculs et établie des statistiques sur un historique donné permettant de valider ou non les hypothèses et d'ensuite les confirmer avec le client.
- Etape 4 : l'actuaire peut réaliser les calculs et ensuite délivrer ceux-ci sous forme d'états comptables qui peuvent être accompagnés ou non d'un rapport d'évaluation pour plus de détails.

Les calculs sont explicités dans la suite de ce chapitre.

1.3.1 Méthodes d'évaluation

Nous parlerons, dans cette partie, des méthodes utilisées afin d'évaluer au mieux la dette actuarielle engendrée par certains avantages au personnel. En effet et pour rappel, les passifs sociaux nécessitant une évaluation actuarielle sont les régimes à prestations définies ou les autres avantages à long terme.

Nous nous intéresserons à la méthode des « Unités de Crédits Projetées » connue sous le nom de « *Projected Unit Credit service prorata* » (*PUC method*) ou encore la méthode « *Projected Unit Credit prorata temporis* » et recommandée par la norme IAS 19 pour le calcul de la dette actuarielle. En effet, lorsque les droits du régime ne sont pas linéaires (on parle de barèmes par paliers) et lorsque les services rendus au cours d'exercices ultérieurs aboutissent à un niveau de droits significativement supérieur à celui des exercices antérieurs (on parle souvent de « plafond » de droits, les convention collective francaises sont pour la majorité d'entre elles plafonnées). Cette méthode instaure pour chaque salarié, la détermination d'une indemnité de départ en retraite calculée à partir de son salaire estimé à la retraite (projeté) mais aussi des droits potentiels en fin de carrière déterminés à l'aide de son ancienneté et du barème de droits prévus par la convention collective (ou par l'accord d'entreprise). L'indemnité théorique est ensuite actualisée et probabilisée. Ce montant va ensuite être proratisé pour correspondre à la période pendant laquelle le salarié a effectivement rendu des services à l'entreprise. En effet, la méthode PUC consiste à répartir l'engagement sur la durée de vie active du salarié au prorata de son ancienneté ce qui implique que l'engagement ou la dette actuarielle correspond à la part de l'obligation allouée aux services rendus à la date de clôture des comptes.

Ainsi pour résumé, pour chaque salarié, on projette le jour de son départ en retraite théorique afin de calculer l'indemnité à laquelle il aurait droit, en prenant en compte les hypothèses de turn-over, de mortalité et d'évolutions des salaires permettant de probabiliser l'engagement. On actualise ensuite ce montant afin de le récupérer en valeur dite à date.

La Valeur Actuelle des Prestations Futures (VAPF) représente la valeur actuelle de l'obligation et est généralement exprimée selon la formule suivante :

$$VAPF = \frac{Prestation\ estimée \cdot Probabilité\ de\ présence}{(1 + Taux\ d'actualisation)^{durée\ résiduelle}}$$

Elle représente la somme actualisée, à la date d'établissement du bilan, des prestations probables. Cette valeur est calculée individu par individu puis sommée.

Par probabilité de présence, on entend probabilité de turn-over (probabilité qu'un individu soit présent dans le régime et donc dans l'entreprise au moment du départ en retraite) et probabilité de mortalité (probabilité qu'un individu soit vivant au moment de la retraite).

La prestation estimée est quant à elle calculée à l'aide de la formule suivante :

$$Prestation\ estimée = \frac{SAB}{12} \cdot (1 + \rho)^{durée\ résiduelle} \cdot Q$$

Avec :

- SAB : Salaire annuel brut du salarié à la date d'évaluation,
- ρ : Taux d'augmentation des salaires ,

— Q : Quotité de droits acquis à la date d'évaluation, exprimée en nombre de mois,

La *Projected Benefit Obligation* (PBO) correspond à la dette actuarielle de l'entreprise à la date d'évaluation. Elle désigne la valeur actualisée des paiements futurs permettant d'effacer l'intégralité de la dette au cours de l'exercice en cours ainsi que des exercices antérieurs. Il s'agit, plus simplement, d'une partie de la VAPF représentant la part acquise de l'engagement à la date d'évaluation qui se traduit par la formule suivante :

$$PBO = VAPF \cdot \frac{Ancienneté\ actuelle}{Ancienneté\ au\ terme}$$

Le *Service Cost* (SC), aussi appelé Charge Normale, représente l'accroissement de la valeur actuelle de l'obligation résultant des services rendus au cours de la période en cours. Autrement dit, il s'agit de l'acquisition d'une année supplémentaire de droit, d'où la terminologie d'unité de crédit. Ceci se traduit comme suit :

$$SC = PBO \cdot \frac{1}{Ancienneté\ actuelle}$$

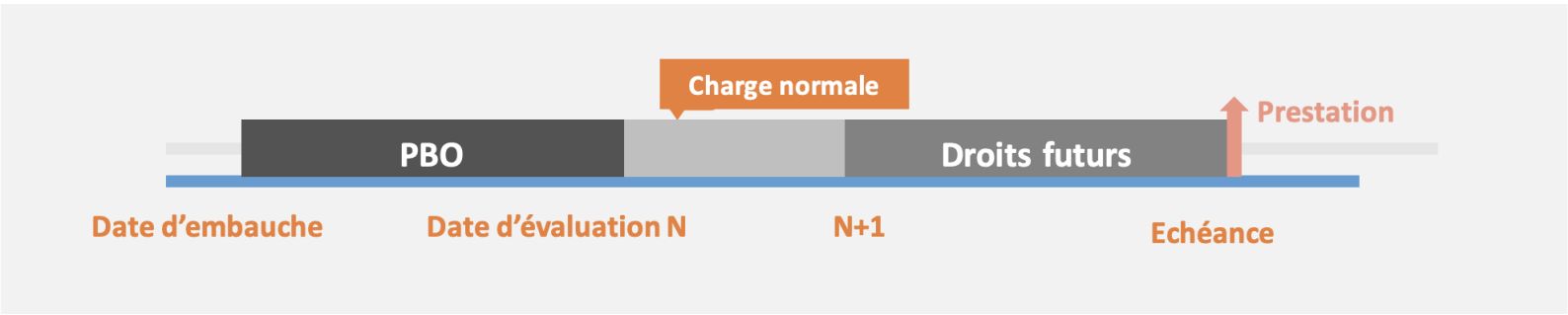


FIGURE 1.2 – Relation entre l'engagement et le service cost

L'*Interest Cost* (IC) ou la Charge d'Intérêt désigne la variation de la valeur actuelle de l'obligation résultant du passage du temps. Il s'interprète de la sorte : plus on se rapproche d'une année de l'échéance de la prestation, soit de la date de versement, plus le facteur d'actualisation de la VAPF diminue, il en résulte une augmentation de la VAPF mesurée par l'Interest Cost. On parle également du cout du temps définit comme ci-dessous :

$$IC = (PBO + SC) \cdot Taux\ d'actualisation$$

Nous pouvons alors calculer la Dette Attendue ou la PBO projetée du prochain exercice qui se calcule avec les éléments précédents de la façon suivante :

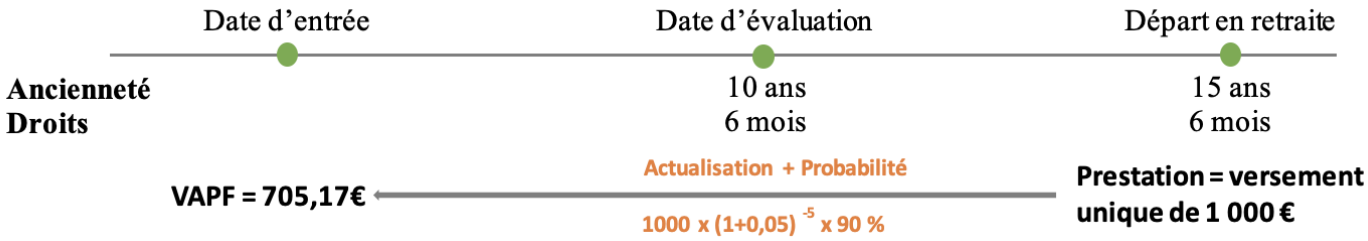
$$PBO_{n+1} = PBO_n + SC + IC - Prestations$$

Cette relation ne donne pas exactement la valeur de la PBO du prochain exercice mais nous fournit plutôt la valeur théorique attendue. En effet, les hypothèses utilisées (taux d'actualisation, évolution de salaire, table de mortalité, table de turnover…) d'une année sur l'autre peuvent différer, ceci créant des écarts actuariels. Nous verrons dans la suite de quoi il s'agit.

Il existe également la méthode dite de « PUC stricte » ou « *Projected Unit Credit acquisition prorata* ». Cette méthode diffère de la méthode PUC au prorata de l'ancienneté puisque cette fois-ci on applique un prorata non plus sur l'ancienneté au terme mais sur l'ancienneté « plafond » des droits. Pour bien cerner la différence, nous allons procéder aux calculs de la PBO pour un même individu selon ces deux méthodes.

Application analytique : Nous prenons le cas d'un barème de droits simplifié présentant un seul et unique palier. Ce barème, fictif, alloue six mois d'indemnité de fin de carrière si un salarié justifie d'une ancienneté supérieure ou égale à 10 ans dans la société et zéro sinon.

Nous calculons la PBO pour un salarié qui totalise dix ans d'ancienneté à la date d'évaluation ayant ainsi atteint le « plafond » de droits. Concernant les hypothèses de probabilité et d'actualisation, nous posons celles-ci égales à 5% pour le taux d'actualisation et 90% pour la probabilité de paiement de la prestation, elles ne sont pas représentatives mais uniquement illustratives.



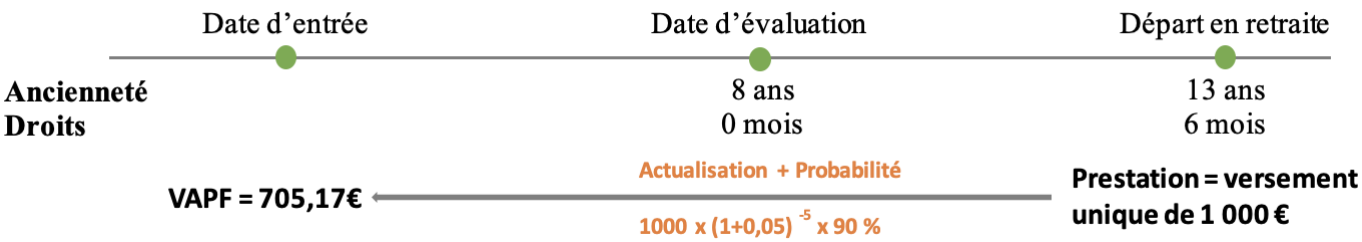
Selon la méthode « PUC service prorata » nous obtenons la PBO suivante :

$$PBO = VAPF \times \frac{Ancienneté\ actuelle}{Ancienneté\ au\ terme} = 705,17 \times \frac{10}{15} = 470,12$$

Alors qu'avec la méthode « PUC acquisition prorata », nous obtenons :

$$PBO = VAPF \times \frac{Ancienneté\ actuelle}{Ancienneté\ plafond} = 705,17 \times \frac{10}{10} = 705,17$$

Nous effectuons le même calcul pour un salarié cumulant huit années d'ancienneté à la date d'arrêté, n'ayant ainsi pas atteint l'ancienneté « plafond ». Nous admettons que celui-ci peut prétendre à une indemnité identique que le salarié précédent et nous calculons l'engagement selon les mêmes hypothèses.



Selon la méthode « PUC service prorata » nous obtenons la PBO suivante :

$$PBO = VAPF \times \frac{Ancienneté\ actuelle}{Ancienneté\ au\ terme} = 705,17 \times \frac{8}{13} = 433,95$$

Alors qu'avec la méthode « PUC acquisition prorata », nous obtenons :

$$PBO = VAPF \times \frac{Ancienneté\ actuelle}{Ancienneté\ plafond} = 705,17 \times \frac{8}{10} = 564,14$$

On constate que, pour un même montant de la valeur actuelle des prestations futures, la méthode « PUC acquisition prorata », alloue une dette actuarielle à date plus importante que la méthode « PUC service prorata ».

A travers notre exemple de barème par paliers, nous observons l'intérêt de cette méthode qui est plus prudentielle que l'approche « service prorata ». Admettons que le premier individu liquide sa retraite avant la date de départ théorique, l'employeur a provisionné trop peu avec la méthode « service prorata » contrairement à la méthode « acquisition prorata ».

Dans ce mémoire, nous nous limiterons pourtant à la méthode « PUC service prorata », largement plus répandue.

1.3.2 Hypothèses actuarielles

Dans cette section, nous précisons chacune des hypothèses intervenant dans le calcul. Pour rappel, l'évaluation actuarielle d'un engagement suppose la projection probabiliste de certains éléments comme la population potentiellement bénéficiaire (en appliquant des probabilités de mortalité et de turn-over via la probabilité de présence), les prestations et le coût de ces dernières au moment du versement. Ainsi plusieurs paramètres sont à intégrer lors de la réalisation des calculs actuariels des engagements sociaux.

« Les hypothèses actuarielles doivent être impartiales et mutuellement compatibles. Les hypothèses actuarielles sont les meilleures estimations faites par l'entité des variables qui détermineront le coût final des avantages postérieurs à l'emploi. » (paragraphes 75 et 76, IAS 19R)

Certaines de ces hypothèses sont entièrement propres à l'entreprise, c'est le cas pour le turn-over et l'augmentation de salaire, par exemple, qui dépendant de la politique interne de la société. Tandis que d'autres sont économiques comme le taux d'actualisation. Certaines peuvent être imposées par la législation en vigueur, notamment pour les tables de mortalité par exemple ou l'âge de départ en retraite.

Ces hypothèses de calcul peuvent être regroupées sous deux familles : les hypothèses démographiques, décrivant au mieux les paramètres qui touchent à la population de bénéficiaires et les hypothèses économiques décrivant au mieux l'environnement économique à la date d'évaluation.

Les hypothèses démographiques

Les tables de mortalité : Les tables de mortalité, permettant de probabiliser la prestation théorique en fonction de la survie des salariés, sont des tables dites d'expérience obtenues à partir d'observations sur une population donnée.

« Une entité doit déterminer ses hypothèses de mortalité en se référant à sa meilleure estimation de la mortalité des participants au régime pendant et après l'emploi. » (paragraphe 81, IAS19 R)

En France, les tables généralement utilisées par les actuaires pour l'évaluation des indemnités de fin de carrière sont les tables dites instantanées établies par l'INSEE, il convient d'utiliser la dernière table parue car elle reflète le plus la réalité actuelle. Mais ils en existent d'autres, par exemple la législation française impose les tables générationnelles TGH TGF 00-05 pour les engagements viagers. Les tables générationnelles se distinguent des tables instantanées par la séparation des individus par génération (année de naissance). Ces tables de mortalité, se présentent sous la forme ci-dessous :

Âge	l_x
0	l_0
...	...
k	l_k
...	...

Âge Année	...	N	N+1	...
0	...	l_0^N	l_0^{N+1}	...
...	...	...	...	...
k	...	l_k^N	l_k^{N+1}	...
...	...	...	...	...

TABLE 1.1 – Tables de mortalité du moment (à gauche), générationnelles (à droite)

Elles nous fournissent ainsi le nombre de survivants à l'âge x desquels nous en déduisons par différence le nombre de décès d_x survenus entre l'âge x et $x+t$. Nous pouvons alors obtenir les probabilités de décès de la manière suivante :

$$q_x = \frac{d_x}{l_x} = \frac{l_x - l_{x+t}}{l_x}$$

Les tables de turn-over : Le taux de turn-over, comme pour les tables de mortalités, permet de prendre en compte dans le calcul de l'engagement la probabilité pour le salarié de quitter la société et donc ne pas finir sa carrière au sein de la société.

On parle de table de turn-over puisque l'usage est d'utiliser des taux de rotation du personnel exprimés en fonction de l'âge. Ces tables sont construites sur la base de statistiques prospectives de départs. Elles peuvent être globales et donc appliquées à tous les salariés de l'entreprise indépendamment de leurs catégories socioprofessionnelles mais peuvent différer selon celles-ci.

Ces tables nous fournissent les probabilités p_x de sortie de l'entreprise à l'âge x et peuvent être par paliers ou décroître linéairement en fonction de l'âge.

L'âge de départ à la retraite : L'âge de départ en retraite intervient lors de la projection des droits des salariés et du calcul de leurs prestations théoriques. Cette hypothèse doit correspondre à la tendance observée au sein de l'entreprise. Il peut être fixe (par exemple 64 ans toute catégorie confondue) mais peut différer selon les catégories socioprofessionnelles.

L'âge de départ en retraite au sens des indemnités de fin de carrière et engendrant le règlement d'une prestation par l'employeur est généralement lié à l'âge de liquidation de la

retraite du régime général et donc l'obtention du taux plein. Il est alors possible d'intégrer ceci au calcul des indemnités de fin de carrière. L'âge de départ est alors individualisé selon les salariés et est fonction de l'année de naissance de ceux-ci. Cette hypothèse de taux plein engendre également la mise en place d'une hypothèse d'âge de début d'activité professionnelle afin d'estimer le point de départ du calcul de la durée d'assurance, cet âge de début d'activité est généralement différencié selon la catégorie socioprofessionnelle.

Les hypothèses économiques

Le taux d'actualisation : Pour rappel, le taux d'actualisation intervient lors du calcul de la valeur actuelle des prestations futures. Contrairement à un taux d'intérêt appliqué sur le marché, le taux d'actualisation est un principe mathématique qui permet d'évaluer à la date d'évaluation des flux qui seront versés dans le futur.

La norme préconise toutefois l'utilisation d'un taux qui reflète le marché. Il convient alors de se fixer un indice de référence représentatif du marché des obligations de première catégorie comme le stipule le paragrpahe ci-dessous :

« Le taux utilisé pour actualiser les obligations au titre des avantages postérieurs à l'emploi (financés et non financés) est déterminé en fonction des rendements du marché à la fin de la période de reporting sur les obligations de sociétés de haute qualité. » (paragraphe 83, IAS 19 R)

En pratique, l'actuaire utilise alors l'indice iBoxx Corporate Bonds de notation AA et de maturité 10+ pour les indemnités de fin de carrière et de maturité 5 et 7 ans pour les médailles du travail, la maturité devant être identique à celle de l'engagement. La notation des obligations des entreprises de première catégorie (AA) est selon l'indice iBoxx une moyenne des 3 notations délivrées par Fitch, Moodys et S&P.

On pourra se référer également : soit à la courbe de taux des obligations à taux fixe, soit au taux unique correspondant au taux moyen des emprunts des obligations d'Etats (OAT) de maturités égales à la durée résiduelle moyenne d'activité du régime.

Le taux de rendement des actifs : Le taux de rendement des actifs à utiliser doit, depuis la norme IAS 19 Révisée, rentrée en vigueur au 1er janvier 2013, être identique au taux d'actualisation. A noter que cette hypothèse intervient uniquement lorsque l'entreprise dispose d'un actif de couverture permettant l'externalisation du financement du régime. Le taux de rendement attendu doit traduire l'évolution de la juste valeur des actifs du régime détenus au cours de l'exercice et doit être en accord avec le marché.

A noter que la comptabilité française autorise toujours le recours à un taux de rendement attendu des actifs différents du taux d'actualisation.

Le taux de profil de carrière : Aussi appelé taux d'augmentation des salaires, il permet de projeter le salaire à la date de départ en retraite du salarié. Il est défini en concertation avec les ressources humaines, l'usage veut que celui-ci se présente sous la forme d'un pourcentage annuel pouvant inclure ou exclure l'inflation qui sera dans ce cas, ajouter séparément ou non. On peut également distinguer un taux d'augmentation des salaires en fonction de l'âge ou

de la catégorie socio-professionnelle, le tout étant que ce taux reflète la politique salariale de l'entreprise à moyen et long terme.

« Les estimations des futures augmentations de salaire tiennent compte de l'inflation, de l'ancienneté, de la promotion et d'autres facteurs pertinents, tels que l'offre et la demande sur le marché de l'emploi. » (paragraphe 90, IAS 19 R)

1.4 Comptabilisation

Les coûts liés aux avantages du personnel accordés en échange de services sont comptabilisés dans les états comptables de la période durant laquelle ces derniers sont rendus. En effet, les engagements liés aux services rendus du personnel ont un impact à la fois dans le résultat net de l'entreprise, au bilan mais également dans les comptes de résultat. Dans cette partie, nous allons définir chaque élément de comptabilité permettant la reconnaissance des engagements au titres des passifs sociaux.

1.4.1 La provision

Tout d'abord, nous explicitons la provision qui désigne le montant à provisionner afin de couvrir financièrement les coûts des passifs sociaux et qui est défini de la manière suivante :

PBO réelle évaluée à la date de la clôture (ex : 31/12/N) - Valeur des actifs de couverture à la date de clôture
= Provision à la date de clôture

FIGURE 1.3 – Les différents éléments de la provision

1.4.2 La charge

Concernant la charge annuelle nette de retraite, celle-ci constitue le coût net sur l'exercice engendré par les engagements sociaux de l'entreprise. Elle est constituées des principales composantes ci-dessous :

Charge normale de l'exercice + Charge financière de l'exercice - Rendement attendu sur les actifs de couverture (au taux d'actualisation) + Impact intégral du coût des services passés (modification de régime) + Impact des évènements spéciaux (réduction /liquidation de régime)
= Charge annuelle de retraite à la date de clôture

FIGURE 1.4 – Les différents éléments de la charge anuelle de retraite

1.4.3 Reconnaissance des écarts actuariels

Pour rappel, de la dette actuarielle nous pouvons calculer le montant théorique de la dette attendue du prochain exercice par l'équation introduite dans la section 1.3.1. Les écarts actuariels proviennent de la différence entre ce montant théorique estimé en $N-1$ pour l'exercice suivant (donc pour l'exercice N durant lequel les écarts actuariels apparaissent) et le montant réel de la dette obtenue lors de l'évaluation N. En effet, ces estimations sont calculées à partir des hypothèses retenues l'année de l'évaluation, l'année suivante l'engagement est recalculé avec une démographie à jour et potentiellement des hypothèses nouvelles.

« Les écarts actuariels sont les variations de la valeur actualisée de l'obligation au titre des prestations définies résultant :
(a) des ajustements liés à l'expérience (l'effet des écarts entre les hypothèses actuarielles antérieures et ce qui s'est effectivement produit) ;
(b) de l'effet des changements apportés aux hypothèses actuarielles. » (paragraphe 8, *Definitions relating to defined benefit cost*, IAS 19 R)

Ceci met donc en évidence deux types d'écarts actuariels, les écarts actuariels dits d'expérience qui surviennent lorsque les hypothèses $N-1$ diffèrent de ce qui est réellement observé entre deux exercices consécutifs (évolution des salaires supérieure ou inférieure à l'hypothèse, turn-over plus important ou plus faible qu'attendu avec les tables de turn-over utilisées) ; nous distinguons également les écarts actuariels liés aux changements d'hypothèses qui comme leur nom l'indique apparaissent lorsqu'un changement d'hypothèses est effectué en N.

On peut donc conclure que les hypothèses d'évaluation conditionnent l'existence de ces écarts actuariels. A ce titre, ces écarts sont à la fois générés sur l'engagement mais également sur les actifs de couverture, il est tout à fait possible de constater un rendement inférieur à celui attendu de par l'utilisation d'un taux de rendement des actifs supérieur à la réalité.

Alors que les écarts actuariels portés sur les actifs de couverture sont uniquement liés au taux de rendement (seule hypothèse utilisée dans leur estimation), leurs identification et justification est alors évidente. Cette tâche peut néanmoins s'avérer plus complexe pour l'engagement du fait du grand nombre d'hypothèses posées pour l'évaluation. Généralement, il faut décomposer l'engagement attendu pour saisir l'apparition des écarts actuariels.

Exemple : Nous nous intéressons au procédé à mettre en œuvre afin de justifier de l'apparition de ces écarts actuariels, et pour cela nous allons utiliser l'exemple simple suivant :

Évaluation N	**Évaluation N+1**
Hypothèses : Turn-over 2% de démissions Évolution des salaires 1,5%	Réel : 500 démissions 100 départs en retraite 1 600 entrées + 2,5% d'évolution des salaires en moyenne
10 000 salariés	**11 000 salariés**

Les écarts actuariels d'expérience constatés sur l'exercice $N+1$ sont :

— environ 3% de gain actuariel d'expérience lié à un turn-over supérieur à ce qui était attendu
— environ 1% de perte actuarielle d'expérience liée à une évolution des salaires supérieures à ce qui était prévu
— une perte liée aux entrées sur l'exercice qui comptabilise d'ores et déjà un engagement minime
— éventuellement des écarts entre les prestations attendues et les prestations réelles

On parle d'une perte actuarielle lorsque les écarts actuariels augmentent la provision. A contrario, lorsque les écarts actuariels réduisent la provision, on parle d'un gain actuariel.

Pour ce qui est de la comptabilisation des écarts actuariels, la norme IAS19 Révisée vient simplifier les méthodes de reconnaissance des écarts actuarielles existantes. En effet, en ce qui concerne les avantages postérieurs à l'emploi l'entreprise comptabilise ceux-ci au bilan via la méthode détaillée ci-après.

Méthode OCI

La méthode *Other Comprehensive Income* (OCI) impacte directement les capitaux propres de l'entreprise. Ainsi, une entité comptabilise les écarts actuariels dès leur apparition, en dehors du compte de résultat. Le montant des écarts actuariels doit être présenté dans un tableau spécifique. Ils sont reconnus et enregistrés dans un compte distinct des autres éléments du résultat global (OCI). Il s'agit de la méthode préconisée dans le référentiel IFRS que nous ne retrouvons pas dans le référentiel français, elle doit être appliquée de manière permanente pour tous les exercices comptables futurs.

A noter qu'en ce qui concerne le référentiel francais, celui-ci laisse la possibilité entre les deux méthodes explicitées, à titre informatif, ci-dessous.

Méthode de comptabilisation directe en résultat

L'entreprise adoptant la méthode de comptabilisation directe intègre immédiatement la totalité des écarts actuariels générés lors de l'exercice dans la charge annuelle nette de retraite et ce, de façon systématique et permanente d'un exercice à l'autre.

Méthode du corridor

L'entreprise ayant fait le choix de la méthode du corridor a la possibilité d'amortir une partie de ces écarts actuariels. Cette fraction d'écarts actuariels est calculée en fonction du stock des gains et pertes actuariels non encore reconnus de début d'exercice, lorsque celui-ci excède ce que l'on appelle la limite du corridor alors l'entreprise comptabilisera des écarts actuariels sur l'exercice. Les limites du corridor sont déterminées comme suit :

$$Limite\ du\ corridor = max(10\% \times PBO_{N-1}, 10\% \times Actifs_{N-1})$$

Le montant de l'amortissement directement comptabilisé est alors calculé en soustrayant à notre stock d'écarts actuariels de début d'exercice le montant de la limite du corridor, puis

en divisant le montant alors obtenu par la durée résiduelle moyenne probable. Le nouveau stock est quant à lui calculé par différence entre le stock de début d'exercice et le montant de l'amortissement.

1.5 Les évènements spéciaux

Au cours de la vie d'un régime, des évènements spéciaux peuvent intervenir impactant alors les différents éléments de la charge ou de la provision de la société. Dans cette partie, nous les explicitons en donnant la définition ainsi que l'impact qu'ils engendrent sur l'engagement.

1.5.1 Modification de régime

Une modification de régime intervient lorsque l'entité change les prestations à payer en vertu d'un régime existant. Cela peut ainsi prendre la forme d'un changement du barème de droit pour les indemnités de fin de carrière ou encore de l'ajout d'une échéance pour les médailles du travail, par exemple. Une modification de régime a un impact (à la hausse ou à la baisse) sur le montant l'engagement.

1.5.2 Réduction de régime

Une réduction de régime intervient lorsqu'une entité peut démontrer qu'elle s'est engagée à réduire de façon significative le nombre de personnes bénéficiant d'un régime, ou change les termes d'un régime de sorte qu'une partie significative des services futurs des employés ne leur donnera plus de droits ou ne leur donnera que des droits réduits.

Exemples : Fermeture d'une usine, abandon de certaines activités, résiliation ou suspension d'un régime, etc…

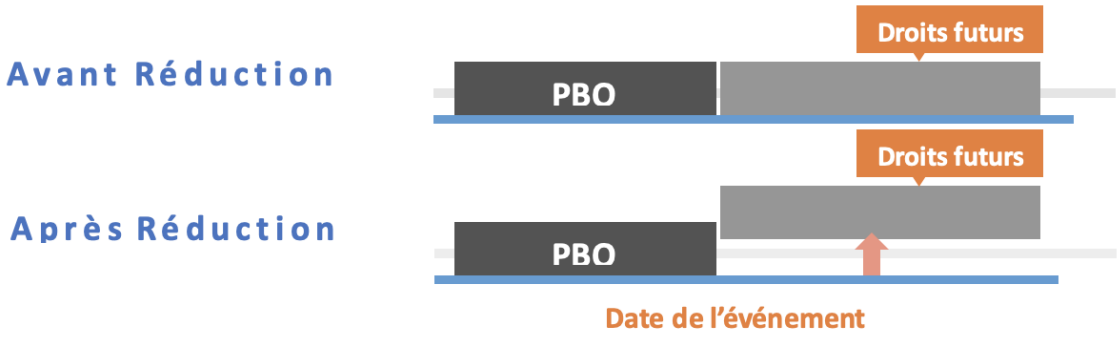


FIGURE 1.5 – Impact d'une réduction de régime sur l'engagement d'une société

Une réduction de régime est un événement susceptible d'avoir des répercussions sur l'engagement liée aux services futurs.

1.5.3 Liquidation de régime

Une liquidation survient lorsqu'une entité conclut une transaction qui supprime toute obligation ultérieure juridique ou implicite relative à tout ou partie des prestations fournies aux termes d'un régime à prestations définies pour des services passés.

Exemples :

- Les participants à un régime reçoivent un montant forfaitaire en espèces en échange de leurs droits à des prestations de retraite déterminées,
- Souscription d'une rente,
- Transfert des obligations à un autre employeur

Schématiquement, nous pouvons dire qu'une liquidation est un événement susceptible d'avoir des répercussions sur l'engagement liée aux services passés.

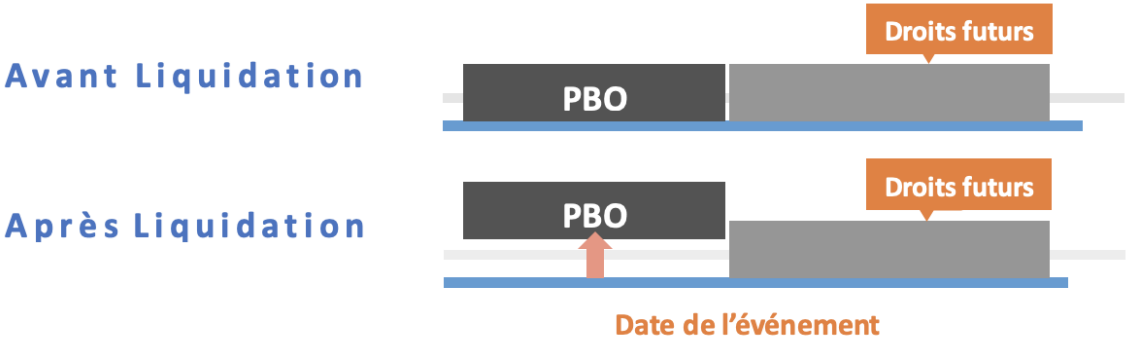


FIGURE 1.6 – Impact d'une liquidation de régime sur l'engagement d'une société

1.5.4 Transferts

Un transfert survient lorsqu'une entité décide de faire fusionner des régimes ou de transférer des employés et leurs obligations vers une autre entité.

1.6 Conclusion du chapitre

Dans ce chapitre, nous sommes revenu sur la norme IAS19. La comptabilisation des passifs sociaux se fait sur des hypothèses démographiques difficilement appréhendable comme le turn-over ou l'évolution des salaires et les enjeux financiers sont importants surtout dans le contexte que l'on connait actuellement où les taux sont à des niveaux très bas entrainant l'augmentation mécanique de l'engagement.

Nous rappelons les élements constituants l'évolution de l'engagement dans le schéma ci-dessous :

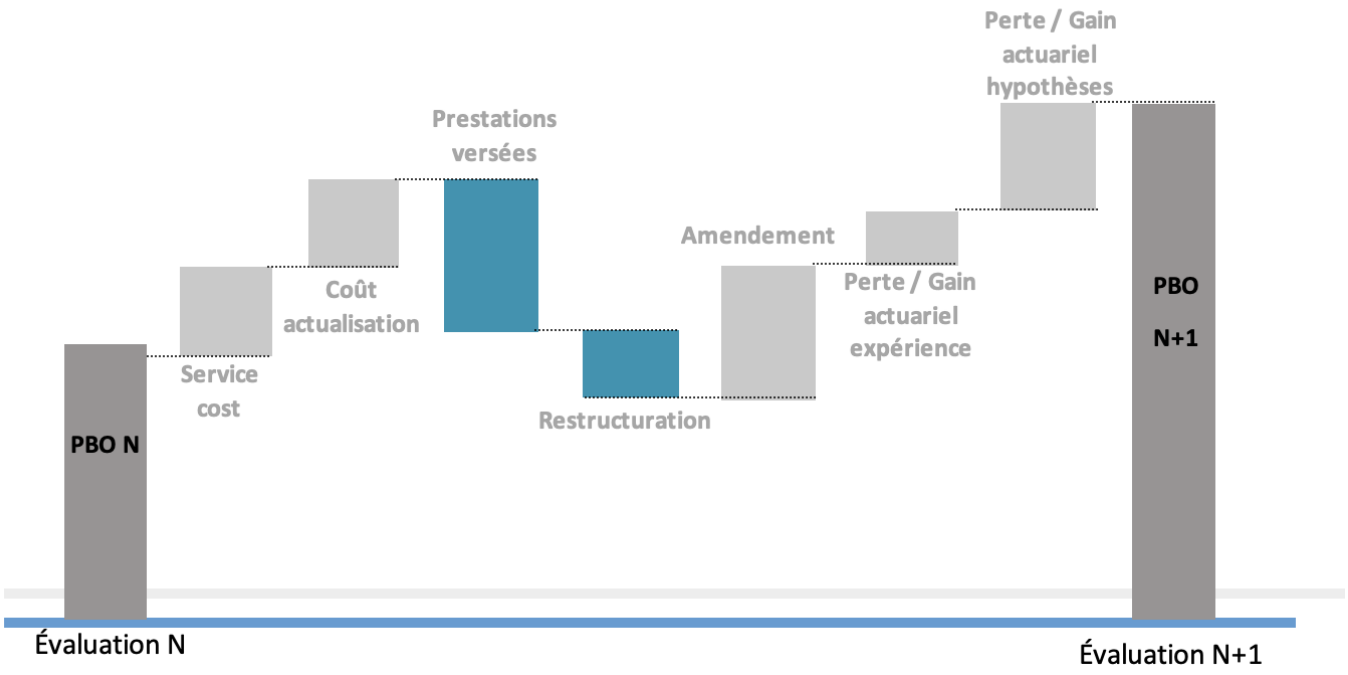


FIGURE 1.7 – Récapitulatif de l'évolution de l'engagement d'une année à l'autre

Chapitre 2

Problématique

Dans ce chapitre, nous commencerons par exposer notre problématique ainsi que le but de cette étude, puis nous expliciterons la méthodologie mise en œuvre afin de constituer notre base de données. Enfin, nous finirons par une première analyse descriptive de celle-ci et une étude de sensibilité sur nos différentes hypothèses.

2.1 L'enjeu de l'étude

Comme nous l'avons vu dans la partie précédente, la norme IAS 19 ne fait pas partie des normes les plus récentes et les procédés permettant le calcul des engagements sociaux sont, à ce titre, bien établis et plutôt stables chez les actuaires.

Le coût potentiel des passifs sociaux n'est toutefois pas à négliger et est susceptible de varier considérablement en fonction des hypothèses de calculs. Il est d'ailleurs assez courant lors du transfert d'un dossier d'un cabinet à un autre de constater des écarts de méthodes et d'hypothèses.

Les hypothèses usuellement appliquées aux calculs ont été explicitées dans le chapitre précédent, nous ne reviendrons donc pas dessus, nous rappelons uniquement celles-ci dans le tableau ci-dessous :

Paramètres et hypothèses : engagement IFC

Date de l'évaluation : 31/12/N

		Type hypothèses
Hypothèses actuarielles		
Taux annuel d'actualisation financière :	0,75%	*Financière*
Taux annuel d'inflation :	1,50%	*Financière*
	Cadres / Etam / Ouvriers	
Age de début d'activité professionnelle :	24 ans / 22 ans / 20 ans	*Démographique*
Age de liquidation des retraites :	Age taux plein	*Démographique*
Evolution annuelle des salaires * :	2%	*Financière*
Tables de survie :	INSEE H & F	*Démographique*
Turn-over :	Tables décroissante par âge s'annulant à partir de 60 ans	*Démographique*
Paramètres IFC		
Scénario de départ :	Départ volontaire	
Attribution des droits :	CCN XX	
Taux de charges sociales patronales :	50%	*Financière*

** inflation comprise*

FIGURE 2.1 – Récapitulatif des hypothèses utilisées en pratique dans le calcul des indemnités de fin de carrière

A noter que les chiffres sont donnés à titre d'exemple uniquement.

Ces hypothèses peuvent parfois se révéler être « simplistes ». En effet, une hypothèse de taux annuel d'évolutions des salaires à $x\%$ ne permet pas de prendre en compte les promotions, ou bien l'effet de l'âge sur l'évolution des salaires. De même, concernant les taux de turn-over, bien que les statistiques empiriques montrent que le turn-over est décroissant avec l'âge, ceci ne permet pas de prendre en compte le niveau de la rémunération ou l'ancienneté par exemple.

Nous allons donc chercher à optimiser les différentes hypothèses de calcul des passifs sociaux car nous souhaitons prendre en compte différents facteurs qui influencent ce calcul mais qui sont habituellement exclus de la pratique. Nous souhaitons, de par cette optimisation, mieux cerner les coûts réels et futurs des passifs sociaux et plus précisément des passifs induits par les indemnités de fin de carrière, puisque nous nous focalisons sur cet avantage au personnel.

2.2 Contexte législatif des indemnités de fin de carrière

Notre choix se tourne vers les régimes d'indemnités de fin de carrière puisqu'il s'agit des régimes les plus fréquents au sein des clients de Secoia et pour cause ils sont obligatoires en France.

En effet, l'article 6 de l'accord annexé à la loi n° 78-49 du 19 janvier 1978 relative à la mensualisation définit l'obligation dite des indemnités de fin de carrière. Cet article a été ensuite repris dans l'article D1237-2 du code du travail.

L'employeur a donc l'obligation de verser des indemnités de fin de carrière à ses salariés lors de leurs départs à la retraite. Cette obligation, induite par la cessation de l'activité professionnelle, est définie par les conventions collectives ou l'accord d'entreprise mais ne peut être

inférieure à l'indemnité légale de départ en retraite si le départ est à l'initiative du salarié ou à l'indemnité légale de licenciement, s'il s'agit d'un départ à l'initiative de l'employeur.

A noter que, dans la suite, nous supposons les départs en retraite comme étant des départs volontaires (à l'initiative du salarié).

Le montant de l'indemnité est calculé à l'aide d'un barème de droit qui dépend de l'ancienneté ainsi que du niveau de rémunération du salarié. Il peut également exister pour une même convention collective deux barèmes de droits en fonction de la catégorie socio-professionnelle.

Avant d'intégrer une optimisation des hypothèses, nous devons tout d'abord construire notre base de données sur laquelle se baseront nos modèles. La suite de ce chapitre traite de la construction du jeu de données.

2.3 RGPD

Avant d'entrer plus en détails quant à la constitution des données, nous rappelons dans cette partie les normes réglementaires relatives aux données personnelles. En effet, les dispositions relatives à la loi « informatique et libertés » du 6 janvier 1978 et au Règlement européen n°2016/679/UE du 27 avril 2016 imposent une règlementation vis-à-vis des données.

En conséquence, aucune données nominatives permettant l'identification des personnes et en particulier les noms, prénoms ne figure dans le présent mémoire.

De plus, dans un souci de confidentialité, aucun client de Secoia ne sera expressément nommé, figurera donc une codification des différentes sociétés sélectionnées dans cette étude.

2.4 Constitution de la base de données

Dans un premier temps, on s'attèle donc à la construction d'une base de données permettant l'élaboration de nos différents modèles explicités dans la suite de ce mémoire. Il s'avère que le cabinet Secoia, implanté à Lyon depuis 1985 et plus récemment à Paris, est spécialisé dans le domaine des évaluations d'engagements sociaux et compte ainsi un certain nombre de clients fidèles depuis plus de 15 ans. C'est ainsi que, basé sur l'historique acquis à l'aide de ces clients, la base de données servant dans ce mémoire a été constituée.

Un inventaire des clients ainsi concernés par un historique de 15 ans a été effectué, il en ressort un certain nombre d'entreprises allant de grands groupes français jusqu'à la simple PME.

Le tableau ci-dessous précise la constitution de l'échantillon :

Dans ce qui suit, nous avons conservé les entités présentes tout au long des 15 exercices.

Échantillon de départ 2005	31
Société sortie du périmètre client	6
Fusion	1
Cessation	-
Liquidation	1
Information indisponible	11
Échantillon final 2019	12

TABLE 2.1 – Constitution de l'échantillon pour la construction de la base de données

2.5 Données retenues et statistiques descriptives

Pour les 12 entités retenues, nous avons regroupé les informations démographiques nécessaires pour l'évaluation des indemnités de fin de carrière composées des éléments ci-dessous :

- **Année :** Année d'exercice (2005 à 2019 dans notre cas),
- **Société :** Codification des différentes sociétés du périmètre,
- **Établissement :** Code postal des établissements de la société,
- **Matricule :** Identifiants du salarié,
- **SEXE :** Code indiquant le sexe du salarié (1 = Homme ; 2 = Femme),
- **DDN :** Date de naissance du salarié,
- **DDA :** Date d'ancienneté du salarié,
- **CSP :** Code indiquant la catégorie socio-professionnelle du salarié à la date de clôture (1 = Cadre),
- **SAB :** Salaire annuel brut de référence pour les Indemnités de fin de carrière,
- **Tx :** Taux d'activité,
- **Âge :** âge du salarié à la date de clôture des comptes,
- **Anc :** ancienneté du salarié à la date de clôture des comptes.

Seuls sont pris en compte dans le calcul de l'engagement, les salariés titulaires d'un contrat de travail à durée indéterminée. Ainsi les salariés en CDD, en contrat de qualification, en contrat d'apprentissage, en cumul emploi-retraite sont exclus de notre base de données. A contrario, les salariés en congé parental, en maternité, en longue maladie, en invalidité, sont, quant à eux, maintenus dans notre base (sur la base d'un salaire théorique identique à celui qu'ils percevraient en activité).

Cet historique, permet également de répertorier les entrées et sorties d'effectif ainsi que les évolutions de salaires au fil des années. Nous disposons donc également d'une base de sortie regroupant les motifs de sorties associés à chaque départ.

La base de données ainsi obtenue est composée d'environ 150 000 lignes. Le traitement des données se fait sous le logiciel R permettant de facilement manipuler une quantité importante de données.

Les statistiques fournies ci-après ont pour objectif de donner quelques points de repères quant à la pertinence de nos données.

Dans un premier temps, nous construisons plusieurs histogrammes permettant de visualiser notre effectif global au cours des années :

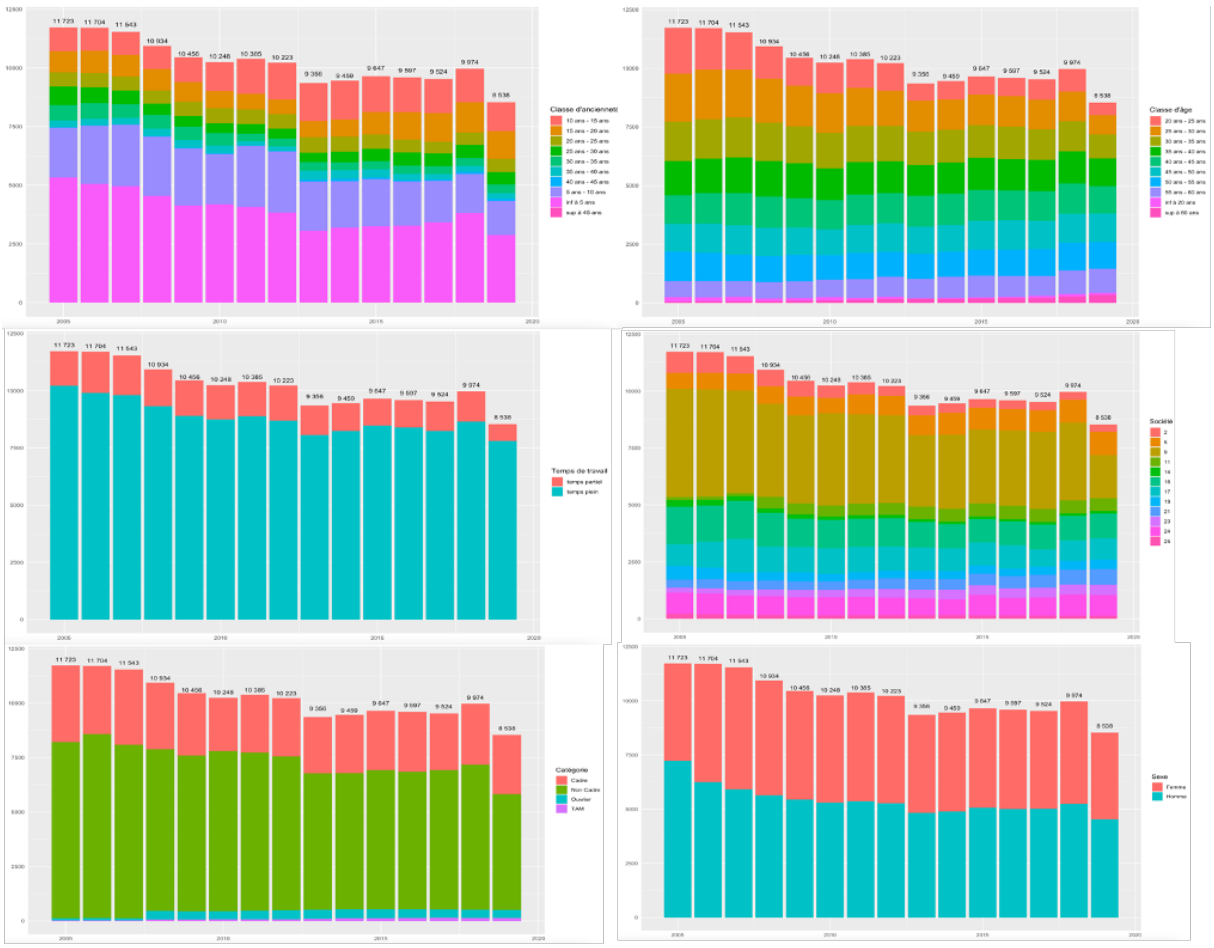


FIGURE 2.2 – Histogrammes de notre effectifs en fonction des années d'exercice

Globalement, notre effectif diminue durant ces quinze exercices. Nous remarquons que les hommes qui représentent 53% de notre effectif global ainsi que les salariés ne bénéficiant pas du statut cadre représentant 68% de l'effectif global sont les populations les plus nombreuses. Les salariés ayant moins de dix ans d'ancienneté sont largement majoritaires et ce durant l'intégralité des quinze exercices. Ceci est confirmé par la pyramide des anciennetés présentées ci-dessous (à gauche).

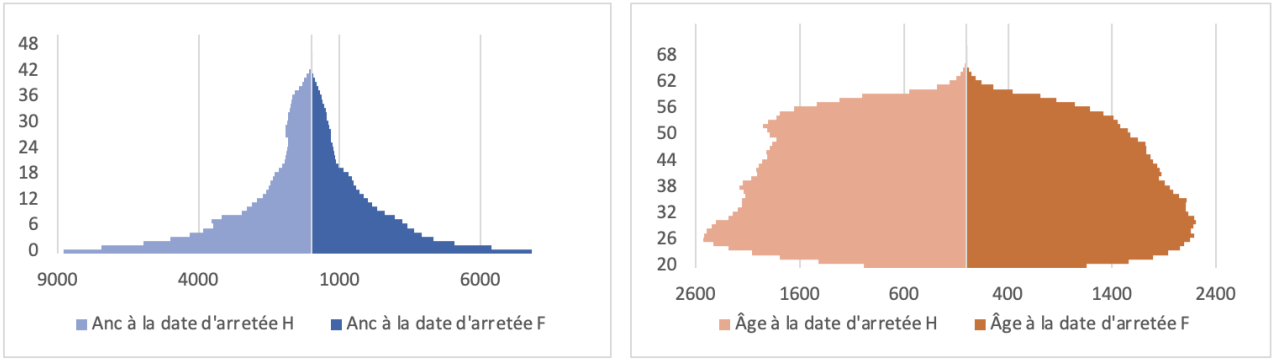


FIGURE 2.3 – Pyramide des anciennetés (à gauche) et la pyramide des âges (à droite)

Aussi, d'après la pyramide des âges (à droite ci-dessus), nous constatons que la population est majoritairement jeune (entre 20 ans et 35 ans) ceci indique un taux de renouvellement des salariés assez élevé. Notre attention est alors particulièrement portée sur les entrées et les sorties et nous traçons l'histogramme ci-dessous chiffrant pour chaque exercice l'effectif stable (ni entrant ni sortant), les sorties, les entrées ainsi que les entrées/sorties.

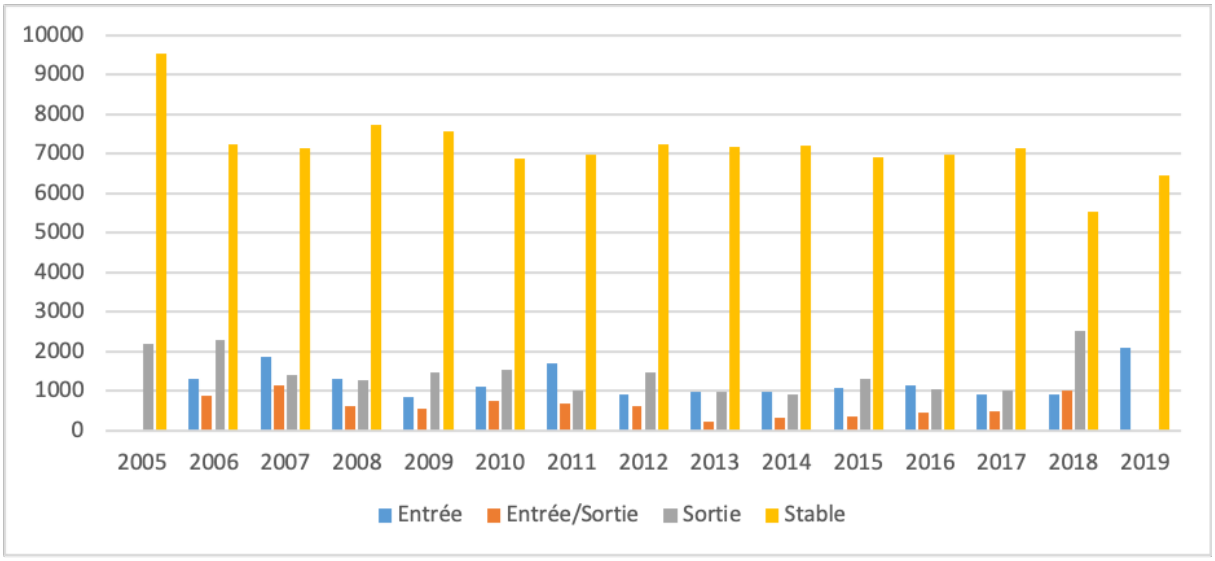


FIGURE 2.4 – Histogramme des mouvements d'effectif en fonction des années

Les entrées/sorties correspondent à des salariés entrés en cours d'exercice puis sortis sur le prochain, avec moins d'un an d'ancienneté, on constate qu'ils sont d'ailleurs assez nombreux puisque peuvent représentent jusqu'à 10% de notre effectif selon l'exercice (notamment en 2018).

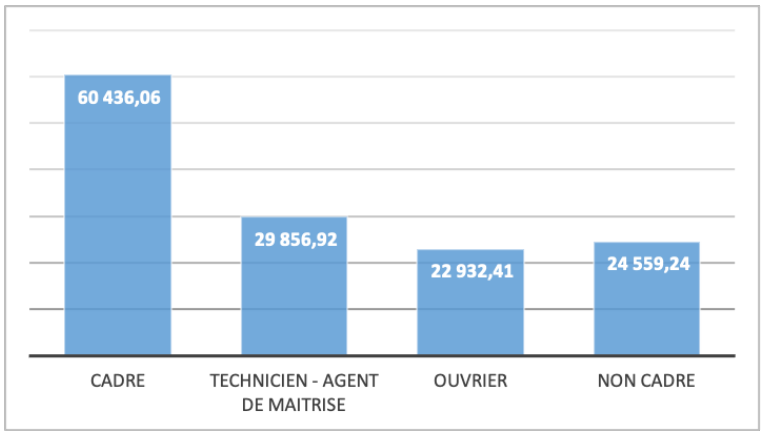


FIGURE 2.5 – Histogramme des salaires moyens selon la catégorie socio-professionnelle

En dernier lieu, nous nous interessons aux salaires moyens de notre portefeuille selon la catégorie socio-professionnelle. Sans surprise, les cadres ont en moyenne un salaire annuel brut plus élevé que les autres catégories socio-professionnelles. A noter que l'appelation Non Cadre comprend à la fois des Techniciens, Agents de Maitrise et Ouvriers, d'où la moyenne plus élevée que les Ouvriers seuls. Ce regroupement a lieu lorsque la Convention Collective ne distingue pas les droits de ces non-cadres et que les hypothèses de calcul ne différent pas d'une catégorie à une autre.

2.6 Sensibilité de l'engagement aux différentes hypothèses

Afin de mener à bien l'étude, un modèle de calcul des indemnités de fin carrière a été implémenté sous le logiciel R. En effet, le logiciel actuellement utilisé par Secoia ne permettait pas de prendre en compte d'autres hypothèses que ce qui a été explicité dans la partie 1 de ce chapitre. Il a donc fallu créer un outil permettant le calcul des indemnités et de prendre en compte le résultat de cette étude, nous reviendrons sur cet outil dans le chapitre 5. L'outil ainsi implémenté est adapté au calcul des indemnités de fin de carrière par la méthode « PUC prorata temporis ».

Nous effectuons dans cette partie une étude de sensibilité de nos hypothèses sur notre engagement (la VAPF) afin de mettre en avant les plus influentes et celles qui devront être optimisées par la suite.

2.6.1 Sensibilité au taux d'actualisation

Pour rappel, le taux d'actualisation permet d'évaluer à la date de clôture (date de calcul des engagements) des flux financiers qui seront versés postérieurement à cette date, dans le futur. Il intervient dans le calcul de la valeur actuelle des prestations futures comme ci-dessous :

$$VAPF = \frac{Prestation\ estimée \cdot Probabilité\ de\ présence}{(1 + Taux\ d'actualisation)^{durée\ résiduelle}}$$

En pratique, nous l'avons vu, le taux d'actualisation utilisé pour le calcul des engagements sociaux est souvent référencé sur le taux des obligations corporate notées AA de maturité 10+. Ainsi l'actuaire utilise un taux unique pour l'ensemble des individus déterminé en fonction de la maturité de l'engagement. Cette méthode est accepté par les auditeurs. Toutefois, l'actuaire peut construire une courbe de taux et appliqué l'approche dite granulaire permettant d'appliquer un taux différencié selon l'échéance de réglement des prestations.

Ci-dessous, nous réalisons une étude de sensibilité en considérant un taux d'actualisation unique pour l'ensemble des salariés, nous présentons le tableau répertoriant les variations de l'engagement en fonction de celles du taux d'actualisation et de la durée résiduelle d'activité du salarié.

Durée résiduelle	+0,5%	+1%	+1,5%	+2%	+2,5%	+3%	+3,5%
0,5	-0,2%	-0,5%	-0,7%	-1%	-1,2%	-1,5%	-1,7%
1,5	-0,8%	-1,6%	-2,3%	-3,1%	-3,8%	-4,6%	-5,3%
3,33	-1,6%	-3,3%	-4,8%	-6,4%	-7,9%	-9,4%	-10,8%
4	-2,0%	-3,9%	-5,8%	-7,6%	-9,4%	-11,2%	-12,9%
…	…	…	…	…	…	…	…
20	-9,5%	-18,0%	-25,8%	-32,7%	-39,0%	-44,6%	-49,7%
34	-15,6%	-28,7%	-39,7%	-49,0%	-56,8%	-63,4%	-68,9%
38,5	-15,9%	-31,8%	-43,2%	-52,3%	-61,4%	-68,2%	-72,7%

TABLE 2.2 – Impact de la variation du taux d'actualisation sur la VAPF

Plus la durée résiduelle est conséquente plus l'impact d'une augmentation du taux d'actualisation augmente considérablement.

2.6.2 Sensibilité au taux de revalorisation des salaires

Le taux d'évolution des salaires permet d'estimer la prestation à laquelle les membres du personnel peuvent prétendre au moment de leur départ en retraite. Ainsi dans les calculs, celui-ci, que l'on va noter ρ, intervient dans la formule suivante :

$$Prestation\ estimée = \frac{SAB}{12} \cdot (1+\rho)^{durée\ résiduelle} \cdot Q$$

Avec, pour rappel, ρ le taux de profil de carrière et Q le quotient de salaire tiré du barème de droits auquel le salarié peut prétendre s'il liquide sa retraite au sein de l'entreprise. En raisonnant de la même façon que pour le taux d'actualisation, nous pouvons en déduire que cette fois-ci plus le taux d'évolution des salaires augmente, plus l'engagement augmente, et cette augmentation dépend également du nombre d'années de service résiduelles avant la liquidation du salarié.

Durée résiduelle	+0,5%	+1%	+1,5%	+2%	+2,5%	+3%	+3,5%
0,5	0,2%	0,5%	0,7%	1%	1,2%	1,5%	1,7%
1,5	0,8%	1,6%	2,4%	3,2%	4%	4,8%	5,6%
3,33	2%	3%	5%	7%	9%	10%	12%
4	2%	4%	6%	8%	10%	13%	15%
…	…	…	…	…	…	…	…
20	10%	22%	35%	49%	64%	81%	99%
34	19%	40%	66%	96%	132%	173%	222%
38,5	20%	48%	76%	116%	160%	212%	276%

TABLE 2.3 – Impact de la variation du taux d'évolution des salaires (inflation comprise) sur la VAPF

Nous pouvons émettre le même constat que pour la sensibilité du taux d'actualisation, plus la durée résiduelle est importante plus l'impact de l'évolution des salaires l'est également.

2.6.3 Sensibilité au taux de mortalité

La mortalité est introduite dans le calcul de la VAPF. Pour mettre en évidence les impacts de celle-ci, nous avons pour une seule même table de mortalité (la table INSEE 2013-2015) augmenté d'une part l'espérance de vie d'un an et d'autre part diminué celle-ci d'une année. Afin d'ajouter une année d'espérance de vie dans la table de mortalité, nous décalons la probabilité de décès d'une année supplémentaire, c'est-à-dire l_{x+1} devient l_x.

Nous constatons que l'impact est symétrique puisque si en augmentant la durée de vie d'un an dans nos tables cela engendre une augmentation de $x\%$ sur notre engagement nous observons une diminution de $-x\%$ dans le cas inverse.

Là encore, l'impact diffère selon la durée résiduelle étant donné que la probabilité totale de survie d'un salarié est calculée sur sa carrière projetée jusqu'au départ en retraite. Cette probabilité est donnée par la formule suivante :

$$p_x = 1 - q_x = 1 - \frac{(l_x - l_{x_{ret}})}{l_x} = \frac{l_{x_{ret}}}{l_x}$$

Où x_{ret} est l'âge du salarié lors de son départ en retraite et x son âge lors de la date de calcul. Ainsi plus l'âge à la retraite du salarié est éloigné de celui à la date de calcul, plus l'impact de la mortalité est important.

2.6.4 Sensibilité au taux de turn-over

Pour rappel, le turn-over intervient directement dans le calcul de la VAPF dans la Probabilité de présence auquel s'ajoute la probabilité de survie. Les sensibilités qui suivent ont été réalisée à l'aide d'une table de turn-over décroissante par âge s'annulant à partir de 60 ans.

Durée résiduelle	+0,5%	+1%	+1,5%	+2%	+2,5%	+3%	+3,5%
0,5	0%	0%	0%	0%	0%	0%	0%
1,5	0%	0%	0%	0%	0%	0%	0%
3,33	-0,01%	-0,03%	-0,03%	-0,05%	-0,06%	-0,07%	-0,08%
4	-0,03%	-0,06%	-0,07%	-0,11%	-0,14%	-0,16%	-0,16%
…	…	…	…	…	…	…	…
20	-3,75%	-7,34%	-10,81%	-14,20%	-17,47%	-20,55%	-23,63%
34	-16,53'%	-30,46%	-42,14%	-51,99%	-60,21%	67,06%	-72,83%
38,5	-23,16%	-41,16%	-55,08%	-65,86%	-74,13%	-80,45%	-85,32%

TABLE 2.4 – Impact de la variation du taux de turn-over sur la VAPF

2.6.5 Sensibilité au taux de charges sociales

Le taux de charges patronales intervient directement dans l'estimation de la prestation de fin de carrière. La variation du taux de charges patronales influe de la même façon pour tous les salariés de l'évaluation. En effet, il va de soi que lorsque ce taux augmente de 1%, toutes les prestations augmentent de 1%.

Le graphique ci-dessous représente la variation sur l'engagement en fonction de la variation du taux de charges patronales.

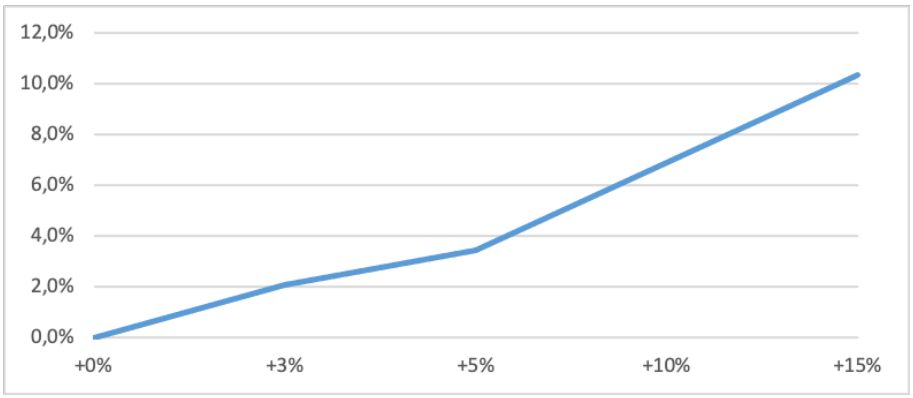


FIGURE 2.6 – Impact de la variation du taux de charges patronales sur la VAPF (utilisation d'un barème par paliers)

A noter que cette sensibilité été réalisée avec un barème de droits par paliers mais reste identique avec l'utilisation d'un barème linéaire, du moins avec l'utilisation de la méthode « PUC prorata temporis ».

2.6.6 Sensibilité à l'âge de départ en retraite

L'âge de départ en retraite intervient à plusieurs niveaux dans le calcul de l'engagement. Supposons dans la suite que nous reculons l'âge de départ en retraite. Le premier niveau d'intervention de cette modification est le calcul de la prestation théorique, un salarié quittant la

société plus tard est susceptible d'acquérir des droits supplémentaires (en fonction du barème car ce n'est peut-être pas le cas lorsque le barème de droits est un barème par paliers ou bien lorsque le barème est plafonné). Ainsi la prestation théorique augmente et on pourrait donc s'attendre à une augmentation de l'engagement.

Toutefois le recul de l'âge de départ en retraite intervient également dans l'allongement de la durée résiduelle et donc dans la probabilisation de cette prestation. Ceci induit une diminution de la probabilité de survie, une augmentation du facteur d'actualisation et n'influe que très peu sur le turn-over. A noter que la probabilité de turn-over est généralement considérée comme nulle passé 60 ans et sont très faibles à partir de 55 ans. On part du principe qu'un salarié proche de la retraite n'a aucun intérêt à quitter la société et perdre les droits qu'il a acquis. En résumé plus un salarié reste en poste au sein d'une société, plus sa probabilité de décès en poste augmente et plus la probabilité de recevoir le versement de sa prestation diminue.

Enfin selon la méthode du prorata temporis, la PBO est aussi impactée puisque le rapport entre l'ancienneté actuelle et l'ancienneté au terme diminue.

2.7 Conclusion du chapitre

A travers ce chapitre, nous avons explicité la constitution de la base de données ainsi que les reconstitutions des effectifs d'une année à l'autre. Enfin, une étude de sensibilité s'est avérée nécessaire afin de cibler les hypothèses à modéliser dans la suite de cette étude.

Nous allons donc nous attacher à construire deux modèles prédictifs l'un concernant le turn-over, l'autre concernant l'évolution des salaires.

Ce choix est, outre le fait que les sensibilités montrent que ces hypothèses influent le plus sur l'engagement, d'autant plus appuyé que :

- pour ce qui est du taux de charges patronales, celui-ci influe peu sur l'engagement et ne constitue pas un critère déterminant,
- pour ce qui est de l'âge de départ en retraite, nous ne disposons pas à ce jour d'assez de recul sur ce que va engendrer la réforme du régime universel des retraites et notamment l'âge pivot, nous préférons donc raisonner sur la base de la legislation en vigueur,
- pour ce qui est du taux d'actualisation, la construction d'une courbe de taux nécessite des données (prix des obligations, etc...) dont nous ne disposons pas. Bien qu'une projection par le modèle de Vacisek pourrait être envisager nous décidons d'écarter cette hypothèse.

Deuxième partie

Modélisations et résultats

Chapitre 3

La modélisation du turnover

Dans cette partie, nous souhaitons modéliser le taux de rotation du personnel plus communément appelé le turn-over. Pour cela nous disposons de notre historique de 15 ans durant lequel les départs et leurs motifs (licenciement, démissions, licenciement économique, etc…) ont été répertoriés. Nous obtenons ainsi le tableau suivant inventoriant les départs par années pour l'intégralité de notre périmètre.

Année	Société	Etablissement	Matricule	Sexe	DDN	DDA	CSP	SAB	Tx	age	anc	Motif de départ
2005	3	74700	4350	1	10/06/1947	04/05/1976	7	12 485.85	0.65	58.6	29.7	1
…	…	…	…	…	…	…	…	…	…	…	…	…

Table 3.1 – Base de données pour la modélisation du turn-over

Les départs sont codifiés par une variable qualitative : 0 non sorti, 1 pour les démissions, 2 départs en retraite, 3 licenciements, etc…

3.1 Les motifs de sortie à intégrer dans notre modélisation

Dans la pratique, la plupart des actuaires se base sur les statistiques historiques de la société pour créer des tables prospectives décroissantes par âge, se pose alors la question de quel motif de sortie intégrer à ces statistiques. La définition légale du turn-over inclue les démissions, les licenciements et les ruptures conventionnelles. Or les commissaires aux comptes en France préconisent de prendre en compte uniquement la démission pour la construction des tables de turn-over comme nous pouvons le voir ci-dessous :

« Dans son étude EC 2018-07, la Commission commune de doctrine comptable de la CNCC et du CSOEC considère que le taux de rotation du personnel à retenir pour l'évaluation des indemnités de fin de carrière (IFC) doit être déterminé en ne tenant compte que des prévisions de démission, à l'exclusion de toute autre hypothèse de départ. » (Commission commune Ordre des Experts comptables et des Commissaires aux comptes, octobre 2018)

Pour comprendre les conclusions ci-dessus, il faut avoir en tête que la législation française prévoit le versement d'une indemnité en cas de licenciement. Aussi, si la société convient avec

un salarié d'une rupture anticipée de son contrat de travail, elle sera tenue de lui verser une indemnité. La démission est donc le seul cas de départ du salarié n'engendrant aucun paiement d'indemnité.

Quant à la norme IAS19, le raisonnement développé selon celle-ci relatif aux modalités de calcul du taux de turn-over, est quelque peu identique à celui préconisé par les commissaires aux comptes français. En effet, la norme IAS 19 impose de provisionner les indemnités de départ dont le paiement est certain en tant qu'avantage postérieur à l'emploi comme le préconise le paragraphe ci-dessous :

« Certains avantages sociaux sont versés quelle que soit la raison du départ de l'employé. Le paiement de ces prestations est certain (sous réserve de toute exigence d'acquisition des droits ou de service minimum) mais le moment de leur paiement est incertain. Bien que ces avantages soient décrits sous certaines juridictions comme des indemnités de départ, il s'agit d'avantages postérieurs à l'emploi plutôt que d'indemnités de cessation de contrat de travail, et une entité les comptabilise comme des avantages postérieurs à l'emploi. » (paragraphe 164, IAS 19 R)

De plus dans le référentiel IFRS, étant donné que les licenciements sont sous le contrôle de l'entreprise, ils ne peuvent être provisionnés qu'à leur date d'annonce. Ainsi, on en conclut que d'inclure les licenciements ou ruptures conventionnelles aux tables de turn-over conduiraient à sous-évaluer l'engagement.

A noter toutefois que le sujet a fait et fait toujours débat et que les avis divergent. Pour certains, le fait d'exclure les licenciements et ruptures conventionnelles reviendrait à provisionner indirectement ces deux motifs à partir d'un barème d'indemnité de départ à la retraite. De plus, les licenciements et ruptures conventionnelles peuvent représenter une part conséquantes des départs constatés au cours d'un exercice (en moyenne sur le périmètre étudié depuis 2005, les licenciements et ruptures conventionnelles représentent 30% des départs constatés) diminuant considérablement les taux de turn-over utilisés jusqu'alors.

Dans ce mémoire, nous excluons les licenciements et les ruptures conventionnelles, c'est une prise de position qui peut être discutée mais en accord avec les recommandations des commissaires aux comptes francais.

3.2 Traitement des données

Nous revenons ci-après sur la manière dont est constituée la base de données servant à la modélisation du turn-over. Il s'agit de récupérer d'une année à l'autre les effectifs de notre périmètre. Ainsi une reconstitution des effectifs est réalisée chaque année mettant en évidence des entrées et des sorties. Avec la validation par nos interlocuteurs des départements ressources humaines, les motifs de sorties sont renseignés.

A la lumière de ce qui est explicité dans la section qui précède, tous les motifs de sortie autre que la démission ne seront pas modélisés. Ainsi d'une année à l'autre, notre effectif se compose d'un effectif stable, c'est-à-dire qui sera présent lors de l'exercice qui suit ainsi que d'un effectif démissionnaire qui va démissionner lors de l'exercice qui suit.

3.3 Analyse de données

3.3.1 Quelques statistiques descriptives

Avant de nous lancer dans la modélisation du turn-over, il convient de mieux cerner les caractéristiques de notre population sortante d'une part, et de notre population dite stable d'autre part (autrement dit les individus qui restent présents dans l'effectif d'une année sur l'autre).

Ci-dessous nous traçons le boxplot de nos différents taux de turn-over par société en fonction des années d'exercice.

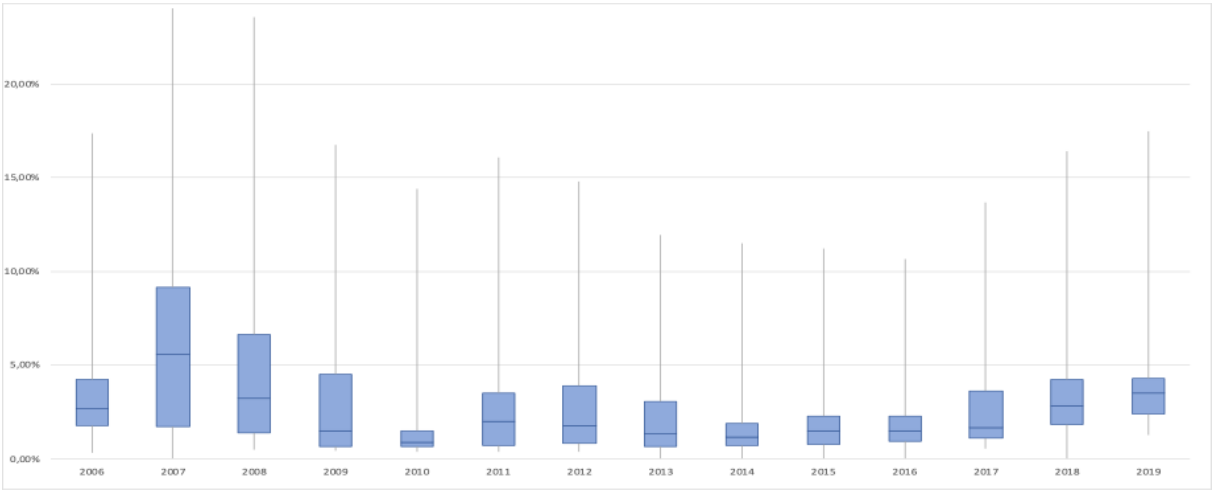


FIGURE 3.1 – Boxplot des taux de turn-over en fonction des années

Nous remarquons que les taux de turn-over varient grandement d'une société à une autre notamment sur les premiers exercices de l'historique étudié. Ceci est à prendre en compte pour la construction du modèle lors de la séparation de nos données puisque la séparation apprentissage/test ne doit pas dénaturer le caractère représentatif de nos échantillons par rapport aux données intégrales. Cet écart-type important entre nos différents taux de turn-over vient des spécificités propres à chacune des sociétés composant notre périmètre.

Nous remarquons notamment que le secteur d'activité de notre périmètre qui enregistre le plus de démissions et ce sur l'historique des 15 exercices est celui du commerce des articles de sports et d'équipements de loisirs. Celui-ci surpasse largement les taux de turn-over enregistrés sur les autres secteurs d'activité. Le second secteur qui enregistre le plus de démissions est celui des hôpitaux privés. Les autres secteurs d'activités ne se distinguent pas particulièrement.

Bien qu'à travers les données dont nous disposons, il est difficile d'appréhender des notions telles que l'ambiance générale au sein de l'entreprise ou la satisfaction au travail des salariés, nous pouvons tout de même de par notre historique, nous faire une idée sur les paramètres influant sur les démissions comme la politique salariale de l'entreprise, le secteur, etc. Le paragraphe qui suit traite justement de ce sujet.

3.3.2 Analyse en composantes principales

Pour cela, il existe plusieurs méthodes d'analyses factorielles (Analyse en Composante Principale, Analyse des Correspondances Multiples, etc..) permettant d'appréhender les différentes relations entre nos données. Le choix du type d'analyse dépend des variables à analyser.

Pour ce qui est de l'Analyse en Composante Principale (ACP), il s'agit d'une méthode d'analyse utilisée afin d'étudier la corrélation entre des variables, des individus ou encore les relations entre variables et individus. L'intérêt d'une ACP survient lorsque notre jeu de données volumineux présente des variables quantitatives, on va alors projeter nos données dans un espace de dimension réduite en définissant des axes de manière à ce que l'information du jeu de données initial soit conservée au maximum.

Dans le cadre de l'ACP, les données sont représentées sous la forme d'une matrice. Ainsi le coefficient x_j^i correspond à la valeur de la $j^{ième}$ variable explicative pour le $i^{ième}$ individu.

Dès lors, nous pouvons représenter chaque individu par le vecteur ligne appartenant à $\mathbb{R}^p$ composé des mesures sur les p variables de notre jeu de données et chaque variable peut être représentée par le vecteur colonne composé des mesures des n individus et appartenant à $\mathbb{R}^n$.

La représentation de nos données est alors difficile puisque ces espaces sont bien souvent de dimension supérieure à 3. L'intérêt de l'ACP prend alors tout son sens puisque va réduire la dimension de l'espace affine en projetant les nuages de points sur des plans ou des droites de manière à le déformer le moins possible.

Les sous-espaces de projections maximisant la distance, ou « l'inertie », sont engendrés par les vecteurs propres de la matrice de corrélation entre les variables explicatives. Ces vecteurs forment les axes factoriels.

Nous précisons que dans notre cas, l'ACP a été centrée afin de rééquilibrer les ordres de grandeur et faciliter l'étude des individus les uns par rapport aux autres et non par rapport à l'origine.

Nous représentons ci-dessous le diagramme des valeurs propres :

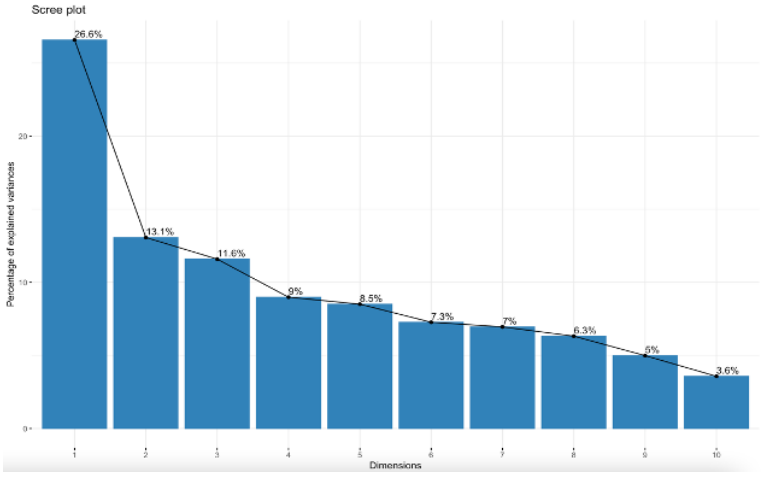


FIGURE 3.2 – Diagramme des valeurs propres

Les valeurs propres renseignent sur la proportion de l'inertie totale prise en compte par chaque axe. On peut donc par une lecture graphique, définir le nombre d'axes à conserver et

sur lesquels seront synthétisées les informations importantes contenues dans notre base de données.

Il existe deux critères de choix résumés dans la table suivante :

Critère de Kaiser	Retenir seulement les axes dont l'inertie est supérieure à l'inertie moyenne qui est égale à 1 dans le cas d'une ACP normée	Nombre d'axes retenus : 3 Valeur d'inertie conservée : 51.2%
Critère de Coude	Retenir seulement les axes avant le décrochement visible sur le diagramme	Nombre d'axes retenus : 1 Valeur d'inertie conservée : 26.6%

TABLE 3.2 – Nombres d'axes à retenir pour l'ACP selon les critères de Kaiser et de Coude

Le choix se porterait sur les trois premières composantes principales à retenir bien qu'elles représentent qu'un peu plus de la moitié de la variance totale, toutefois afin de conserver une quantité plus ou moins conséquente de la variance nous retenons les cinq premières composantes et conservons donc 68.7% de l'inertie totale.

De plus la variable « départ » qui nous intéresse est très bien représentée sur la composante principale 5.

En pratique, on représente le cercle des corrélations qui, à chaque point-variable, associe un point dont la coordonnée sur un axe factoriel est une mesure de la corrélation entre cette variable et le facteur. Nous traçons ce graphique de corrélation des variables ci-dessous :

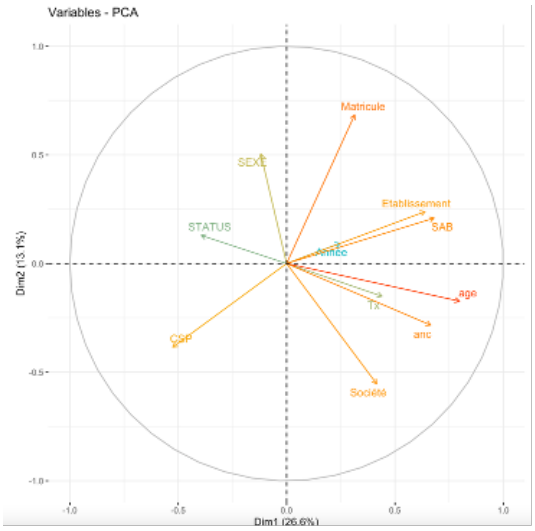


FIGURE 3.3 – Graphique de corrélation des variables sur le plan engendré par les deux premiers axes

Nous constatons que les variables « âge », « ancienneté » et « société » sont positivement corrélées, tout comme le « salaire », le « taux d'activité » et l'« établissement », enfin la « catégorie » semble quant à elle corrélée avec le genre.

Aussi, les départs sont négativement corrélés avec l'âge, l'ancienneté et la société, tout comme la catégorie socio-professionnelle est négativement corrélée avec le salaire. Toutefois

l'aspect « négativement » n'a, dans ce cas, aucun sens puisque l'ACP fonctionnant sur des données numériques, nous avons codifiés nos variables. Il faut ainsi comprendre, par exemple, que plus la catégorie est élevée en codage moindre est le salaire (pour se donner une idée les cadres sont codifiés par 1).

Ici, on effectue notre ACP sur uniquement les variables numériques et que nous traçons un biplot intégrant à la fois la représentation des variables mais également celle des individus :

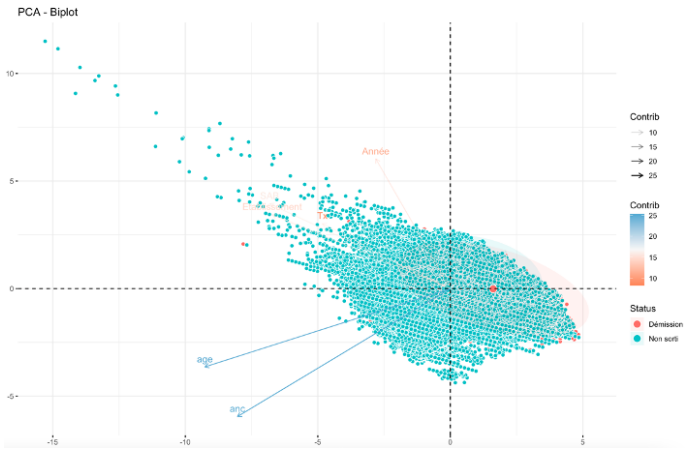


FIGURE 3.4 – Double visualisation des variables et individus sur le plan engendré par les deux premiers axes

Ce biplot nous permet de visualiser simultanément les effets des différentes variables et individus sur nos départs. A noter que la distance entre les individus indique leurs similitudes de comportement vis-à-vis des variables et, on le rappelle, les proximités entre variables indiquent leurs corrélations. Entre autres, nous constatons que plus un salarié est âgé moins il est concerné par les démissions, la flèche « âge » s'éloignant de l'ellipse formée par les individus ayant démissionnés (en rouge). Nous constatons le même effet concernant la variable « salaire » indiquant que plus un salarié évolue au sein d'une entreprise et est « fidélisé » moins il démissionne. Mais tout ceci n'est finalement pas très surprenant. Toutefois le biplot met l'accent sur l'effet de la variable « taux d'activité » qui sera d'ailleurs confirmé par le même biplot sur la composante 3, composante sur laquelle la variable « taux » est la mieux représentée. Nous observons que la flèche représentant le « taux d'activité » s'éloigne de l'ellipse des démissions. Ceci indique que le taux d'activité est un facteur influant dans les démissions, plus un salarié travaille à temps partiel plus il a tendance à démissionner.

Nous venons de voir que le turn-over dépend de la société et donc du secteur d'activité, mais d'autres paramètres sont à prendre en compte comme la catégorie socio-professionnelle, l'ancienneté dans l'entreprise ou encore l'âge. Ces informations vont nous être utiles lors de la sélection des variables de nos différents modèles.

3.4 Modélisation

Le but ici est maintenant de construire un modèle de prédiction des départs dans le cadre de démissions uniquement. Afin de construire un tel modèle, nous allons nous baser sur plu-

sieurs méthodes issues d'une part de l'apprentissage automatique supervisé et d'autre part des modèles linéaire généralisés. La modèlisation consiste alors à prédire une variable à expliquer notée y à partir de variables explicatives notées x à l'aide d'un modèle f tel que $f(x) = y$.

Dans notre cas, on parle de classification puisque y prend des valeurs dans un ensemble fini de classes (ici deux on parle plus précisément de classification binaire). Nous allons plus précisèement utiliser une application de la classification en utilisant le *scoring*. Le *scoring* consiste à générer des valeurs basées sur un modèle d'apprentissage automatique supervisé, à partir de nouvelles données d'entrée. Dans notre cas précis de la modélisation du turn-over, nous souhaitons obtenir la probabilité d'affectation d'une observation à nos deux classes. Ceci vient donc transformer le problème non plus de classification mais en régression.

Le terme supervisé signifie alors que le modèle f et les paramètres à estimer sont déterminés à partir d'observations tirées de notre base de données.

Ainsi, pour revenir sur les notations, nous notons Y l'indicateur caractérisant la situation du salarié lors de l'exercice suivant la clôture des comptes de la société. Y est une variable aléatoire catégorielle prenant deux modalités : 1 (pour « sorti dans le cadre d'une démissions ») et 0 (pour « non sorti »).

Dans la suite de cette partie, nous allons expliciter les différents algorithmes implémentés afin d'en évaluer les performances et de choisir le plus adéquat pour la problématique du turnover.

Ayant ainsi posé les termes de ce problème, nous nous tournons alors naturellement vers d'une part les modèle dits paramétriques des modèles linéaires généralisés et d'autre part les modèles non paramétriques du *machine learning*. Dans la suite, nous revenons sur l'aspect théorique dérrière ces différentes notions.

3.4.1 Les Modèles linéaires généralisés

Généralités

Les modèles linéaires généralisés (GLM) permettent d'étudier la relation entre une variable à expliquer notée $Y = (Y_1, \ldots, Y_n)$ et des variables explicatives notées X_i. Il s'agit d'une généralisation de la régression linéaire simple entre la variable Y et les variables X_i.

Les hypothèses relatives aux régressions linéaires simples, notamment quant à la normalité des résidus qui induisent également certaines hypothèses sur la variable à expliquer, restreignent le champ d'application de ces modèles et c'est bien pour palier à ceci que les modèles linéaires généralisés sont apparus. En effet, la fonction dite lien permet de modéliser des phénomènes bien plus complexes que les régressions linéaires simples.

Les GLM se composent de trois parties :

- la composante aléatoire représentées par la variable à expliquer dont les densités appartiennent à la famille de loi exponentielle (famille de distribution spécifique au GLM),
- la composante déterministe représentée par les prédicteurs linéaires qui forment une combinaison linéaire des variables explicatives connues,
- la fonction lien définie comme une fonction déterministe, strictement monotone sur $\mathbb{R}$.

Contrairement au modèle linéaire classique dans le cas généralisé, ce n'est pas la variable à expliquer qui est directement modélisée mais la fonction lien de l'espérance de cette variable.

$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_p \cdot X_p$$

Définition famille exponentielle : Font partie de cette famille, les lois de probabilité à deux paramètres dont la densité peut d'écrire sous la forme :

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

Où :

- $\theta \in \mathbb{R}$ est le paramètre canonique (ou de la moyenne),
- $\phi \in \mathbb{R}$ est le paramètre de dispersion,
- $a(.)$ est une fonction définie sur les réels et non nulle,
- $b(.)$ est une fonction définie sur les réels et 2 fois dérivables,
- $c(.)$ est une fonction définie sur $\mathbb{R}^{\not\in}$.

Pour une variable aléatoire Y dont la densité s'écrit sous forme exponentielle, nous obtenons :

$$\mathbb{E}(Y) = \frac{\partial b(\theta)}{\partial \theta} = b^{'}(\theta) \qquad \text{et} \qquad Var(Y) = b^{"}(\theta) \cdot a(\phi)$$

L'estimation des paramètres β_i du modèle se fait généralement par maximum de vraisemblance. Reprenons $Y = (Y_1, \ldots, Y_n)$ notre vecteur à expliquer dont la densité f_Y s'écrit sous forme exponentielle :

$$f_Y(y) = exp\left(\frac{\sum_{i=1}^{n} w_i \cdot (y_i\theta_i - b(\theta_i))}{\phi}\right) + \sum_{i=1}^{n}(c_i(y_i, \phi))$$

$\mathbb{E}(Y_i) = b^{'}(\theta_i) \Longleftrightarrow \theta_i = b^{'^{-1}}(\mathbb{E}(Y_i))$, $b^{'}$étant inversible

$$\begin{aligned} & g(\mathbb{E}(Y_i)) = \beta_0 + \beta_1 \cdot X_{1i} + \cdots + \beta_p \cdot X_{pi} \\ \Longleftrightarrow\ & \mathbb{E}(Y_i) = g^{-1}(\beta_0 + \beta_1 \cdot X_{1i} + \cdots + \beta_p \cdot X_{pi}) \text{ , g étant bijective} \end{aligned}$$

Ainsi on peut écrire la fonction de vraisemblance en fonction des β_i comme ci-dessous :

$$\theta_i = b^{'^{-1}}(\mathbb{E}(Y_i)) = b^{'^{-1}} \circ g^{-1}(\beta_0 + \beta_1 \cdot X_{1i} + \cdots + \beta_p \cdot X_{pi})$$

Comme leur nom l'indique les modèles linéaires généralisés présupposent un effet linéaire entre les variables explicatives et la variable à expliquer. Dans la suite nous supposons donc qu'il existe un effet linéaire entre nos démissions et nos différentes variables explicatives.

La régression logistique

On note $y_i = 1$ si le $i^{ième}$ salarié a quitté la société au titre d'une démission $y_i = 0$ sinon. Nous considérons que les individus ont tous la même probabilité de quitter la société, il est alors raisonnable de supposer que les variables aléatoires $Y_1, \ldots, Y_n$ sont indépendantes et de même loi (i.i.d.). Nous verrons dans la suite en quoi cette hypothèse peut être remise en cause. Dans ce cas, on peut supposer que x_i est la réalisation d'une variable aléatoire X_i de loi de Bernoulli de paramètre p et Y suit une loi binomiale de paramètre (n, p) .

Nous nous tournons alors la régression logistique qui est un cas particulier des modèles linéaires généralisés.

Définition : Soit Y une variable à valeurs dans $0, 1$ à expliquer par p variables explicatives $X = (1, X1, \ldots, Xp)$'. Le modèle logistique propose une modélisation de la loi $Y|X = x$ par une loi de Bernoulli de paramètre $p_\beta(x) = \mathbb{P}_\beta(Y = 1|X = x)$ telle que :

$$log\left(\frac{p_\beta(x)}{1 - p_\beta(x)}\right) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_p \cdot x_p = x^{'} \cdot \beta$$

ou encore

$$logit(p_\beta(x)) = x^{'} \cdot \beta$$

logit (ou log-inverse) désigne la fonction dite lien qui est bijective et dérivable de $]0, 1[$ dans $\mathbb{R} : p \mapsto log(\frac{p}{1-p})$

Dans ce cas, on peut écrire :

$$p_\beta(x) = \mathbb{P}_\beta(Y = 1|X = x) = (exp(x^{'} \cdot \beta))/(1 + exp(x^{'} \cdot \beta))$$

La quantité $\frac{p_\beta(x)}{1-p_\beta(x)}$ désigne ce que l'on appelle un odd ratio, un rapport de chance.

La fonction de vraisemblance pour une loi binomiale s'écrit :

$$L(\beta, x) = \prod_{i=1}^{n} p_\beta(xi)^{y(x_i)} \times (1 - p_\beta(xi))^{(1-y(x_i))}$$

La notion de vraisemblance réside dès lors que nous avons un échantillon observé, la vraisemblance quantifie alors la probabilité que les observations proviennent effectivement d'un échantillon théorique de la loi de probabilité associée.

La méthode du maximum de vraisemblance consiste donc à produire les paramètres qui maximisent la vraisemblance par rapport aux données en maximisant la probabilité d'observer cet échantillon. Ceci se traduit par l'annulation de la dérivée du logarithme de la vraisemblance en béta : $\frac{\partial ln(L)}{\partial \beta} = 0$. Une fois la résolution du système d'équation ci-dessous effectué (généralement par l'algorithme de Newton-Raphson), nous pouvons estimer la probabilité $p_\beta(x)$ et même en déduire les odds.

Une fois les paramètres $\hat{\beta}$ estimés, nous calculons la quantité suivante s'appelant la déviance :

$$D = -2 \cdot ln(L)$$

où $ln(L)$ est la log-vraisemblance.

La déviance est contrairement à la log-vraisemblance est positive. Plus la déviance est proche de 0, meilleur est le modèle estimé.

La méthode de régression logistique semble alors plus contraignante en termes d'hypothèses sur la variable à expliquer par rapport aux méthodes dites non-paramétriques qui, nous le verrons, procèdent à l'estimation des probabilités sans hypothèses. Toutefois, on se rendra compte dans la suite que la méthode de la régression logistique reste opérationnelle même lorsque l'on s'écarte assez fortement des hypothèses que la régissent. En effet, on verra qu'il existe réellement un effet linéaire entre nos données et les démissions.

3.4.2 Le Machine learning

Nous nous tournons désormais vers l'apprentissage automatique. L'apprentissage automatique s'inscrit dans le cadre de l'intelligence artificielle basée sur l'idée que les systèmes peuvent apprendre des données, identifier des modèles et prendre des décisions avec une intervention humaine minimale.

Dans la suite de cette section, nous nous concentrons, dans un premier temps, sur les algorithmes d'apprentissage automatique basés sur les arbres de décision, nous les introduisons donc dans la partie qui suit.

Les arbres de décision

L'algorithme CART (*Classification And Regression Tree algorithm*) fut introduit par Breiman en 1986, c'est un algorithme de partitions binaires récursives permettant de construire des modèles de prédiction

Il repose sur la théorie des graphes qui consiste à subdiviser un échantillon de données de manière binaire afin de segmenter la population selon des critères.

Nos données d'origine sont d'abord séparées en deux en fonction d'une variable d'entrée et d'un seuil sur cette variable créant ainsi un nœud autrement appelé condition. Ensuite, en fonction de cette condition les données sont à nouveau séparées en deux en fonction d'un nouveau seuil et ainsi de suite. Plus on descend dans l'arbre, plus on cumule les conditions et plus l'arbre devient spécifique et complexe.

L'algorithme de construction d'un arbre de décisions se fait selon les itérations suivantes :

Algorithm 1: Algorithme CART

Result: Arbre CART
1- Construction de l'arbre maximal T_{max};
Input : échantillon d'apparentissage à p variables;
while *critère d'arrêt non respecté* **do**
  Sélection de la variable j;
  Calcul de la mesure d'impureté du partitionnement selon la variable j;
  if *mesure d'impureté est optimale* **then**
    Partition des observations selon la variable j;
  end
end
2- Élaguage de l'arbre maximal T_{max};
Input : arbre maximal T_{max} obtenu lors de la phase 1- ;
while T_k *différent du noeud racine* **do**
  Construction de l'arbre T_{k-1} arbre T_k élagué ;
  if $err_{T_{k-1}} < err_{T_k}$ **then**
    Retenir T_{k-1} ;
  else
    Retenir T_k;
  end
end

Ainsi l'algorithme débute au nœud racine avec toutes les instances de l'échantillon d'entraînement, puis sélectionne un attribut sur la base de critères de division (soit le rapport de gain ou autres mesures d'impureté) et enfin l'algorithme partitionne les instances en fonction de l'attribut sélectionné de manière récursive.

Pour comprendre, la notion d'impureté, il faut savoir qu'un sous-ensemble de données est pur si toutes les observations appartiennent à la même classe. L'objectif lors de la création d'un arbre de décision est alors de réduire autant que possible les impuretés dans les données.

La métrique utilisée dans CART pour mesurer l'impureté est souvent l'indice de Gini. L'indice de Gini favorise les plus grandes partitions et est calculé pour $i = 1, \ldots, n$ nombre d'attributs (c'est-à-dire variables) :

$$Gini = 1 - \sum (p_i)^2$$

Ici, p_i est la probabilité qu'un individu du nœud courant appartienne à la classe C_i.

L'algorithme CART cherche donc pour chaque nœud supplémentaire à maximiser le gain d'information, mesuré via le critère de Gini introduit ci-dessus, apporté par les variables explicatives candidates.

Le partitionnement cesse lorsque le critère d'arrêt est respecté. Celui-ci peut être défini au choix par l'utilisateur. Soit l'algorithme cesse lorsqu'il n'y a plus de variable explicative servant à partitionner les données, soit lorsque toutes les données d'un nœud appartiennent à la même classe, ou encore lorsque l'utilisateur fixe un nombre minimum d'instances associé à chaque nœud pour poursuivre la subdivision ou enfin un nombre d'instances minimum que doit contenir une feuille (nœud terminal) pour également poursuivre la subdivision.

L'algorithme CART est une méthode non paramétrique, et à ce titre ne nécessitent pas d'hypothèse sur la distribution des données. Cela présente un avantage considérable comparé à l'approche paramétrique des GLM car approximer une distribution pour les données est souvent difficile et peut conduire à des résultats faussés. De nombreux modèles sont non linéaires et sont tout de même traités linéairement par les méthodes GLM pour plus de simplicité car il est difficile de préciser un modèle lorsque les relations ne sont pas linéaires et que le nombre de variables explicatives est important. Ainsi souvent, le « vrai » modèle est approximé par un modèle linéaire, or lorsque l'on veut prédire des données différentes de celles qui ont servi à construire le modèle, les prévisions risquent d'être très mauvaises. Les arbres sont des algorithmes non linéaires ce qui permet de relâcher cette hypothèse forte des GLM.

Les Random Forests

Les forêts aléatoires, ou *Random Forest* en anglais, sont constituées d'un large nombre d'arbres de décisions qui opèrent comme un ensemble. Il s'agit donc d'une agrégation des différents arbres de décisions.

La prédiction du random forest résulte de la moyenne des prédictions de chaque arbre constituant le random forest et donc repose sur le principe qu'un grand nombre d'arbres de décisions relativement non corrélés fonctionnant comme un ensemble surpassent les modèles individuels. Ceci entraine donc une augmentation de la robustesse du modèle tout en conservant les avantages des arbres de décisions en termes de non-paramétrage.

L'algorithme de construction d'une forêt aléatoire repose sur les étapes suivantes :

Algorithm 2: Algorithme Random Forest

Result: Random Forest
Input : échantillon d'apprentissage de taille n à p variables, N nombre d'arbres à construire et $mtry$ le nombre de variables candidates pour le partitionnement;
for k = 1,...,N **do**
  Tirage avec remise d'un échantillon de données de taille j<n;
  Construire un arbre CART T_k sur l'échantillon;
  for *chaque* $noeud_i$ **do**
    Tirage uniforme de $mtry$ variables parmi p pour former la décision associée au $noeud_i$
  end
end

L'algorithme du Random Forest consiste à créer des sous-échantillons au hasard dans le jeu de données d'apprentissage. Ainsi chaque arbre composant la forêt aléatoire est entrainé sur un de ses sous-échantillon accentuant le caractère non corrélé si cher au bon fonctionnement de l'algorithme. Pour diminuer d'autant plus la corrélation entre les modèles et donc diminuer la variance du modèle agrégé, l'algorithme des forêts aléatoires effectue un autre échantillonnage sur les variables. L'algorithme dans lequel seul le tirage des individus est randomisé (appelé bootstrap) est connu sous le nom de Bagging. Les forêts aléatoires ajoutent au Bagging une randomisation sur les variables explicatives, appelé feature sampling. En effet, la faible corrélation entre les modèles est la clé. Les modèles non corrélés peuvent produire des prédictions d'ensemble plus précises que n'importe laquelle des prédictions individuelles. La

raison de cet effet est que les arbres se protègent les uns les autres de leurs erreurs individuelles (tant qu'ils ne se trompent pas tous constamment dans la même direction). Alors que certains arbres peuvent être erronés, de nombreux autres arbres auront raison, de sorte qu'en tant que groupe, les arbres peuvent se déplacer dans la bonne direction.

Le Boosting

Historiquement l'algorithme des arbres de décision, CART est apparu puis s'en est suivi une méthode d'agrégation des arbres de décision nommée *Bagging*, apparue en 1996 par Breiman. Enfin, c'est la méthode du Random Forest qui est apparu et vient améliorer la méthode du *Bagging* notamment en ajoutant l'échantillonnage des variables par bootstrap comme évoqué précédemment.

Dans cette partie, nous explicitons l'algorithme nommé Boosting qui est lui-même une amélioration de l'algorithme Random Forest. Apparu en 1999 par Freund & Shapire, l'idée à l'origine de l'algorithme est de convertir un modèle faible en apprentissage, c'est-à-dire un modèle légèrement meilleur qu'un modèle aléatoire, pour qu'il devienne plus performant en termes de prédiction.

Contrairement aux Random Forests, l'algorithme Boosting ne s'appuie pas sur du boostrap afin de construire différents modèles parallèlement mais agrège de manière séquentielle des modèles construits les uns à la suite des autres en filtrant les observations prédites correctement. Ainsi chaque modèle essaie de corriger son prédécesseur en se focalisant sur les observations mal classées. Cette filtration est faite par pondération des données à chaque itérations l'algorithme alloue plus de poids aux observations mal classées.

Il existe plusieurs algorithmes de Boosting dont l'algorithme nommé Adaboost. Celui-ci se compose de modèles apprenants faibles constitués d'arbres de décision avec une seule division (c-à-d. un seul étage), appelée souches de décision. Lorsque AdaBoost crée sa première souche de décision, toutes les observations sont pondérées de manière égale. A l'itération suivante, pour corriger l'erreur précédente, les observations mal classées ont désormais plus de poids que les observations correctement classées.

L'algorithme Boosting est le suivant :

Algorithm 3: Algorithme Boosting

Result: Boosting
Input : échantillon d'apprentissage de taille n à p variables;
Initialisation des poids w_1 (pour n données d'observations le poids est initialement identique pour toutes les instances soit $\frac{1}{n}$);
for k = 1,...,N **do**
  Construire un arbre CART T_k sur l'échantillon pondéré à l'aide des poids w_k;
  Calculer l'erreur $\epsilon_k = \frac{\sum_{i=1}^{n} w_i \cdot 1(y_i \neq T_k(x_i))}{\sum_{i=1}^{n} w_i}$;
  while *les poids ne convergent pas* **do**
    Calculer le poids de l'itération $\alpha_k = ln\left(\frac{1-\epsilon_k}{\epsilon_k}\right)$;
    Calculer les poids à jour $w_{k+1} = w_k \times exp\{\alpha_k \cdot 1(y_i \neq T_k(x_i))\}$;
    Normaliser les poids pour que la somme fasse 1 ;
  end
end

Tout comme l'algorithme AdaBoost, le Gradient Boosting, un autre cas particuler de l'algorithme Boosting, fonctionne en ajoutant séquentiellement des prédicteurs à un ensemble, chacun corrigeant son prédécesseur. Cependant, au lieu de changer les poids pour chaque observation classifiée de manière incorrecte à chaque itération, la méthode Gradient Boosting essaie d'ajuster le nouveau prédicteur aux erreurs résiduelles commises par le prédicteur précédent.

Il utilise ce que l'on appelle le *Gradient Descent* pour identifier les lacunes des prédictions du modèle précédent. L'algorithme Gradient Boosting peut être donné en suivant les étapes :

- Ajuster un modèle aux données, $F_1(x) = y$
- Ajuster un modèle aux résidus, $h_1(x) = y - F_1(x)$
- Créez un nouveau modèle, $F_2(x) = F_1(x) + h_1(x)$

Enfin, on trouve l'algorithme XGBoost qui signifie eXtreme Gradient Boosting. XGBoost est une implémentation d'arbres de décision boostés par gradient conçus pour la vitesse et les performances. Les algorithmes de Gradient Boosting sont généralement très lents à mettre en œuvre en raison de la formation séquentielle des modèles. Ainsi, XGBoost se concentre sur la vitesse de calcul et les performances du modèle. C'est ce modèle parmi tout les algorithmes de Boosting que nous avons choisis pour la suite notamment de par la vitesse de calcul que celui-ci offre.

3.4.3 L'avant-propos à la construction des modèles

L'échantillonnage

Nous allons tout d'abord procéder à l'échantillonnage de nos données. L'échantillonnage permet d'une part d'estimer les différents paramètres de nos modèles concurrents à partir desquels on construit des règles de prévisions et d'autre part de mesurer la performance de nos

modèles en termes de prévision et d'en estimer, entre autres, les probabilités d'erreur associées à ces règles.

Nous allons, premièrement, séparer de manière arbitraire notre base de données en échantillon d'apprentissage, échantillon test et échantillon de validation. Ainsi nos différents échantillons comprennent :

— les effectifs de notre périmètre de 2005 à 2016 : *échantillon d'apprentissage*,

— les effectifs de 2017 et 2018 : *échantillon test*,

— les effectifs de 2019 : *échantillon de validation*.

Nous le rappelons, notre base de données est constituée des effectifs à la date d'arreté N et la variable à expliquer $Motif\ de\ départ$ renommée dans la suite $STATUS$ indique le fait qu'un individu ait démissionné au cours du prochain exercice et est donc absent dans la base $N+1$. Il nous faut donc deux exercices consécutifs pour établir notre base.

Cet échantillonnage induit probablement un biais dans notre construction de modèle puisque dépend des observations dans nos différents échantillons. Par exemple, des observations aberrantes peuvent être présentes dans l'échantillon test faussant alors les performances évaluées sur celui-ci. L'erreur peut être surestimée ou à l'inverse sous-estimées en n'utilisant qu'une partie des données pour l'estimer.

Ainsi, afin de supprimer tout biais pouvant être engendrer par l'échantillonnage mais également afin d'éviter le phénomène de sur-apprentissage on utilise généralement la validation croisée, introduite dans la section qui suit.

Le phénomène de sur-apprentissage désigne le fait qu'un modèle prédictif produit par un algorithme s'adapte bien à l'échantillon d'apprentissage. Ceci sous-entend qu'elle s'adapte même trop bien aux données d'apprentissage. Par conséquent, le modèle prédictif capturera les corrélations ainsi que le bruit produit par les données ce qui entrainera l'incapacité du modèle à se généraliser sur des données nouvelles telles que l'échantillon test ou de validation, les modèles « sur-entrainé » font preuve de qualité prédictive faible.

Dans la suite ce cette section nous allons expliciter la procédure de construction de nos différents modèles sur l'échantillonnage apprentissage/test afin de simplifier la compréhension de la construction des différents modèles mais les modèles finaux seront construits et tester en utilisant la validation croisée.

Notre choix final de modèle présentant les meilleures performances se fera sur l'échantillon de validation.

La validation croisée

Il existe plusieurs types de validations croisée comme la validation *leave-one-out* qui consiste à exclure à chaque itération une observation de nos données, de calibrer le modèle sans cette observation et d'en mesurer les performances sur cette observation. La validation *leave-one-out* est très couteuse en termes de temps de calcul, nous préférons donc la validation croisée dite *K-fold*.

Le principe de la validation croisée *K-fold* est le suivant : les données sont partitionnées en K sous-ensembles de taille identique. A chaque itérations les échantillons d'apprentissage

et de test sont construits à l'aide de ses sous-ensembles : le modèle est entrainé sur construire l'échantillon d'apprentissage (exclusion faite d'un sous-ensemble) et on prédit sur l'échantillon test (constitué du sous-ensemble exclu de l'apprentissage) pour mesurer le modèle. Ceci est répété plusieurs fois jusqu'à ce que chaque sous-échantillons des données ait servi une fois à la validation du modèle.

Le schéma ci-dessous synthétise ce procédé :

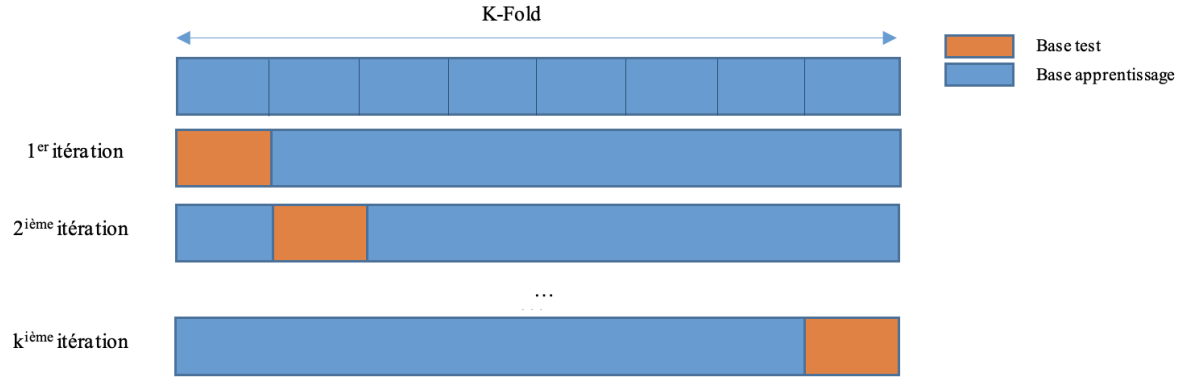


FIGURE 3.5 – Procédé de la validation croisée K-fold

La validation croisée peut également être utiliser afin d'estimer ce que l'on appelle en apprentissage automatique, les hyper-paramètres d'un modèle de prédiction. Ces paramètres concernent les algorithmes d'apprentissage automatique, nous les introduiront lorsque nous aborderons le sujet plus loin dans ce chapitre.

Pour rappel, la manipulation des données a été faite sous R, nous continuerons de travailler sous ce logiciel pour la modélisation.

3.5 La mise en place de la régression logistique

Le modèle implémenté est synthétisé dans le tableau ci-dessous :

Variable à expliquer	Variable binaire ($STATUS$) indiquant le fait de démissionner ou non (pour rappel 0 non sorti et 1 pour démission)
Variables explicatives	Sexe, CSP, ancienneté, âge, salaire annuel brut, taux d'activité, Société, Année
Fonction lien	Fonction logistique
Loi	Binomiale

TABLE 3.3 – Récapitulatif du modèle du turn-over par régression logistique

3.5.1 Création du modèle et sélection des variables

Pour débuter la création de notre modèle, nous nous concentrons sur la sélection des variables puisque seules sont retenues celles qui influent réellement le modèle et améliore celui-ci.

La première approche que nous mettons en œuvre est appelée backward. Elle consiste à créer le modèle le plus complet en termes de variables explicatives et d'enlever à chaque étape la moins significative. L'idée étant de tester l'apport d'une variable donnée en présence de toutes les autres. Nous pouvons observer les résultats du test ci-dessous :

```
> drop1(glm(STATUS ~ Année + Société + SEXE + age + anc + CSP + Tx + SAB, data = MYtrain[c(MYinput, MYtarget)], family = binomial(logit)), test =  "Chisq")
Single term deletions

Model:
STATUS ~ Année + Société + SEXE + age + anc + CSP + Tx + SAB
        Df Deviance   AIC     LRT  Pr(>Chi)
<none>       46755 46797
Année    1    46926 46966  170.67 < 2.2e-16 ***
Société 11    47996 48016 1241.40 < 2.2e-16 ***
SEXE     1    46781 46821   26.25 2.994e-07 ***
age      1    47185 47225  430.41 < 2.2e-16 ***
anc      1    47365 47405  609.83 < 2.2e-16 ***
CSP      3    46785 46821   29.65 1.637e-06 ***
Tx       1    47349 47389  594.45 < 2.2e-16 ***
SAB      1    46764 46804    9.04  0.002634 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 3.6 – Résultat de la fonction *drop1*

Ci-dessus nous présentons les résultats fournis par la fonction drop1 de R, nous permettant de conclure quant à la significativité des coefficients et ainsi la pertinence des variables dans le modèle retenu.

Cette fonction utilise le test du rapport de vraisemblance pour mesurer l'effet d'une variable sur le modèle. Le test de vraisemblance est alors calculé pas à pas et à chaque itération on supprime une variable. La statistique de test est alors :

$$LR = D_S - D_M$$

Avec, $S < M$ et D_S est la déviance du modèle comportant S variables et D_M est la déviance du modèle comportant M variables, les deux modèles étant emboités. Sous H_0, le coefficient associé à la variable supplémentaire dans M est nul, la statistique de test suit une loi du χ^2 de degré de liberté $ddl = ddl_S - ddl_M$.

Ainsi, d'après ce qui précède, toutes les variables sont significatives au seuil de 0.002 au sens du test du rapport de vraisemblance, la suppression de la moins significative SAB n'améliore pas significativement le modèle. Nous arrêtons donc la procédure de retrait de variables ici.

Nous nous tournons ensuite vers l'approche duale dite forward, le contraire de l'approche backward, qui consiste à partir du modèle le plus simple (modèle comprenant uniquement la constante) d'ajouter pas à pas des variables en testant à chaque fois la significativité de la réduction de déviance associée. Les résultats sont visibles ci-dessous :

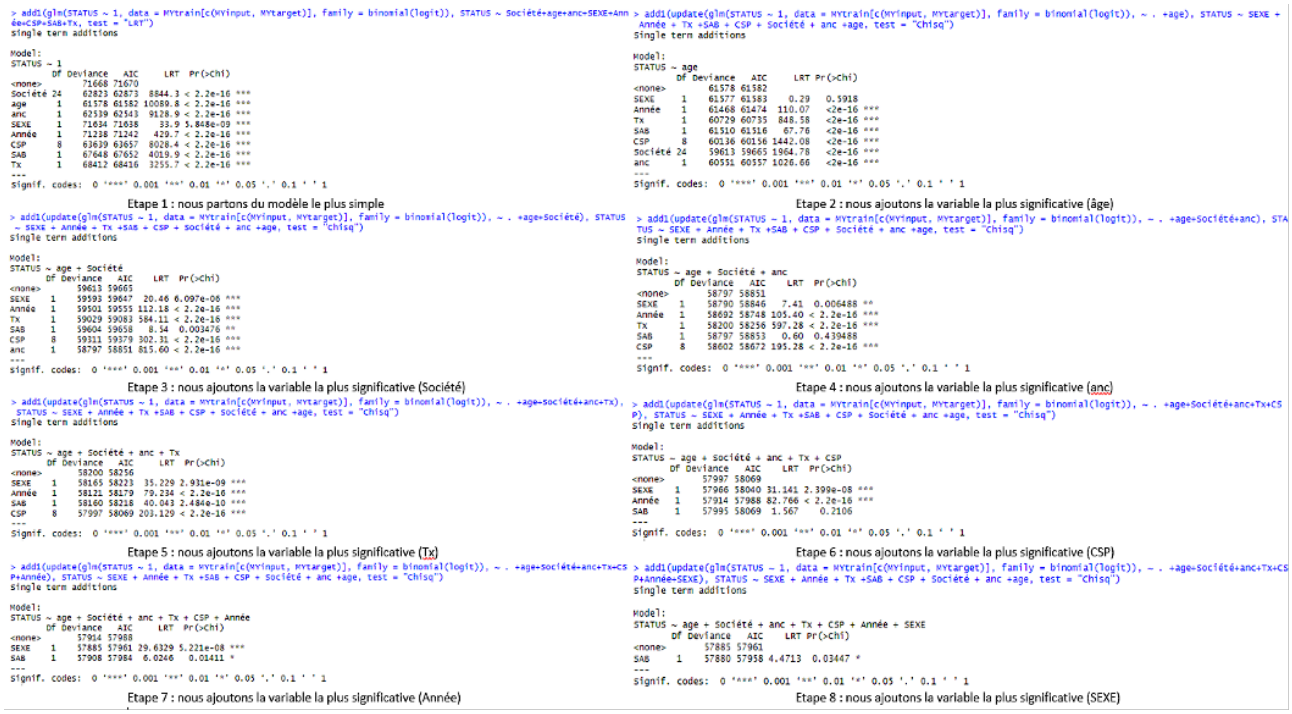


FIGURE 3.7 – Résultat de la fonction *add1*

Ces deux approches, nous donne donc le même modèle ce qui nous permet de juger de la qualité de celui-ci en testant le pouvoir explicatif des variables. Toutefois, nous émettons un doute sur l'apport de la variable SAB dans le modèle puisque d'après ce qui précède il s'agit de la variable la moins significative.

Nous allons donc créer deux modèles l'un incluant le plus grand nombre de variables explicatives, l'autre sans la variable SAB qui semble n'apporter que très peu d'information au modèle.

Nous nous tournons alors vers le critère de l'Akaike Information Criterion afin de comparer les deux modèles en question. Le critère AIC est définit comme suit où L est la vraisemblance et p est le nombre de paramètres du modèle :

$$AIC = -2 \cdot ln(L(\beta)) + 2 \cdot p$$

Le critère AIC est une mesure de qualité qui permet de palier le fait que plus un modèle est complexe plus la vraisemblance augmente. Ainsi, si on se base uniquement sur la vraisemblance on aurait tendance à sélectionner le modèle comportant le plus de paramètres.

Nous utilisons la fonction *stepAIC* de R. Les résultats de cette fonction sont présentés ci-dessous :

```
Step:  AIC=46803.92
STATUS ~ age + Société + anc + Tx + Année + CSP + SEXE

Step:  AIC=46796.87
STATUS ~ age + Société + anc + Tx + Année + CSP + SEXE + SAB
```

Nous sélectionnons le modèle qui présente le critère AIC le plus faible. Ainsi quel que soit le critère le modèle à priori optimum inclut la totalité des variables explicatives de notre base.

A ce stade, nous pouvons émettre l'hypothèse d'interactions entre différentes variables comme l'âge et l'ancienneté par exemple qui évoluent de la même façon d'un exercice au suivant pour un même individu. Toutefois, il existe surement d'autres phénomènes dits d'interaction pouvant être intégrés dans le modèle. Dans la suite, nous allons intégrer ces phénomènes par la construction de modèles concurrents intégrant plusieurs interactions. La méthodologie utilisée est assez simple puisque consiste en l'ajout d'interaction entre variables dans notre GLM et de tester la significativité de cet ajout via le test de Wald et le rapport de vraisemblance comme pour le choix des variables à interaction d'ordre 0 précédemment. A noter que nous procédons de façon à que notre modèle soit « hiérarchiquement bien formulé », ainsi pour une interaction d'ordre n notre modèle doit contenir toutes les interactions d'ordre inférieures pour les variables concernées.

Cette approche s'est avérée non concluante. En effet, l'ajout d'interactions n'améliore pas significativement la qualité du modèle et ne réduit pas significativement la déviance. De plus, pour des raisons de simplification et afin d'éviter la saturation de notre modèle par un sur-paramétrage, nous excluons dans la suite les interactions. Le modèle retenu est donc le suivant :

```
STATUS ~ age + Société + anc + Tx + Année + CSP + SEXE + SAB
```

Nous présentons ci-dessous les odds ratio du modèle ainsi obtenu dans le tableau suivant :

Variables	OR	95% CI	p-value
Anc	0.91	0.91, 0.92	<0.001
Année	0.95	0.95, 0.96	<0.001
Age	0.95	0.95, 0.96	<0.001
Tx	0.21	0.18, 0.23	<0.001
Société			
2	—	—	
5	7.24	5.11, 10.6	<0.001
9	11.2	8.02, 16.4	<0.001
11	4.39	1.85, 9.26	<0.001
14	2.54	1.37, 4.55	0.002
16	3.33	2.34, 4.92	<0.001
17	6.28	4.44, 9.21	<0.001
19	4.30	2.87, 6.59	<0.001
21	2.44	1.64, 3.73	<0.001
23	3.10	2.07, 4.75	<0.001
24	1.79	1.18, 2.76	0.007
25	7.09	4.67, 11.0	<0.001
CSP			
1	—	—	
3	0.67	0.29, 1.71	0.4
4	0.56	0.28, 1.29	0.13
5	0.81	0.75, 0.88	<0.001
Sexe			
1	—	—	
2	0.88	0.83, 0.92	<0.001
SAB	1.00	1.00, 1.00	0.001

TABLE 3.4 – Odds Ratio et Intervalles de confiance associés à la régression logistique entrainée sur l'*échantillon d'apprentissage*

A noter que le logiciel R prend par défaut la première modalité comme modalité de référence,

nous n'avons pas modifié cette modalité de référence puisque ne fausse pas les résultats du modèle. Cela influe simplement sur la manière dont sont interprétés les résultats. D'après le tableau qui précède, nous comprenons par exemple :

- Une femme (SEXE=2) a 0.88 fois moins de chances qu'un homme de démissionner,
- Un salarié non-cadre (CSP=5) a 0.81 fois de chance en moins qu'un cadre de démissionner,
- Un salarié qui acquière un an d'ancienneté supplémentaire a 0.91 de chance en moins de démissionner,
- Un salarié qui acquière un an supplémentaire a 0.95 de chance en moins de démissionner,
- On constate qu'avec un odd ratio à 1, la variable SAB n'a à priori pas d'impact dans les départs en démission.

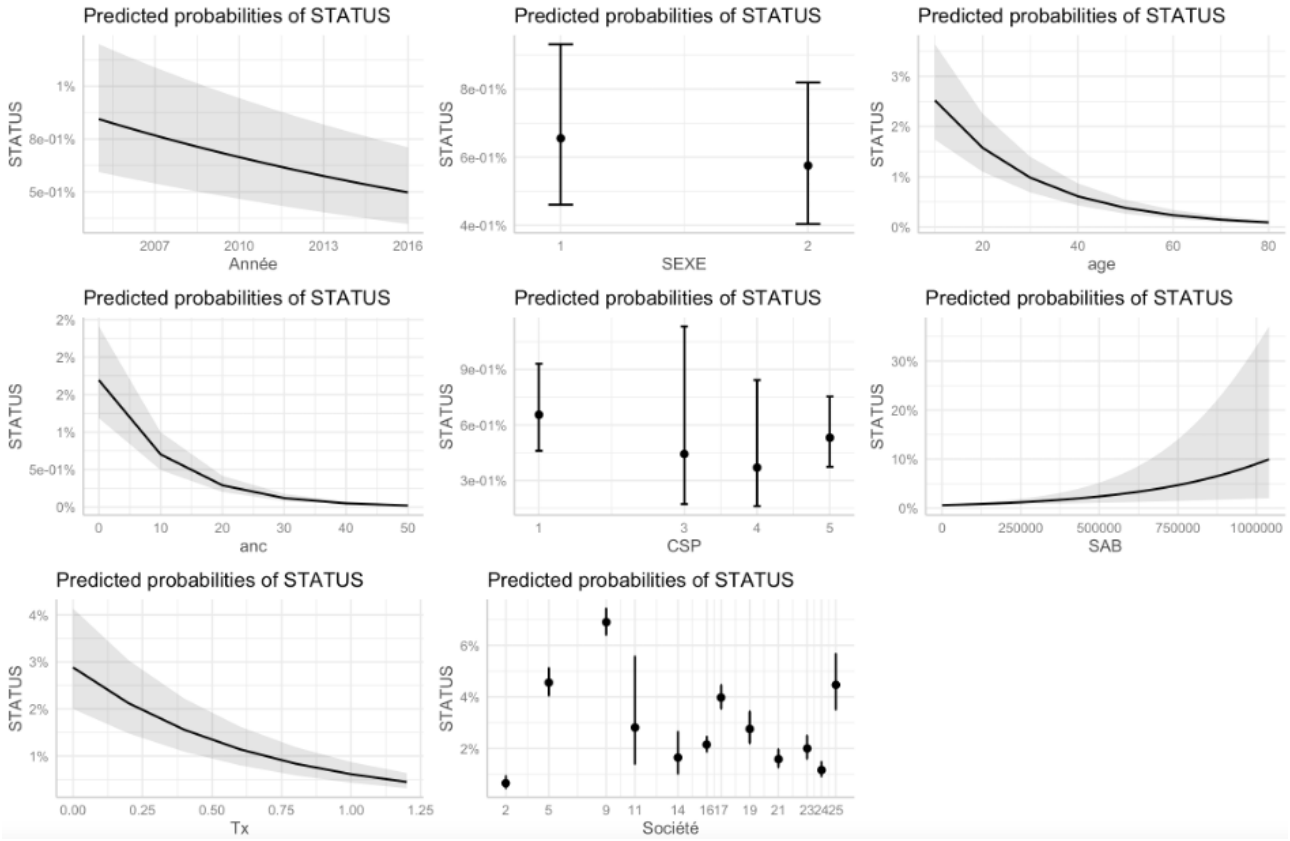


FIGURE 3.8 – Probabilités prédites à l'aide de la régression logistique entrainée sur l'*échantillon d'apprentissage* en fonction des variables explicatives

3.5.2 Mesure de la qualité d'ajustement du modèle

Plusieurs outils nous donnent une idée de la qualité d'ajustement d'un modèle, parmi lesquels nous retrouvons la statistique de Pearson, la statistique de la déviance et le pseudo R^2 de Mc Fadden, etc… Ces outils nous permettent d'apprécier la qualité d'ajustement de notre modèle aux données observées (autrement dit à notre échantillon d'apprentissage).

Pour valider un modèle linéaire généralisés on peut se pencher sur les résidus de Pearson et de Déviance ainsi que leur analyse graphique en représentant ceux-ci en fonction des valeurs réelles. En général, on combine ce graphique par un qqplot afin de vérifier la normalité des résidus. Cependant, ces interprétations ne sont généralement pas valables lorsque le modèle en question est une régression logistique, nous ne nous y attardons donc pas.

Nous calculons tout de même le pseudo R^2 associé au modèle retenu. Le pseudo R^2 est défini comme suit où $ln(L_M)$ est la log-vraisemblance de notre modèle et $ln(L_0)$ celle du modèle dit trivial (uniquement constitué de la constante) :

$$R^2 = 1 - \frac{ln(L_M)}{ln(L_0)}$$

Nous constatons que notre modèle présente un pseudo R^2 de 0.21, ceci peut s'interpréter comme notre modèle améliore de 21% la vraisemblance par rapport au modèle nul et explique 21% de la variance.

Il semble alors indispensable de tester la significative globale du modèle par le test du rapport de la déviance consiste à comparer la déviance du modèle retenu avec celle du modèle trivial :

$$LR = D_0 - D_M$$

Cette statistique de test suit, sous H_0, tous les coefficients sont nuls, une loi du χ^2 de degré de liberté $ddl = ddl_0 - ddl_M$. Au seuil α, le modèle est globalement significatif si $LR > \chi^2_{1-\alpha}$.

On en conclut que le modèle est globalement significatif au seuil 0.005.

Enfin, on confronte les probabilités estimées par le modèle et celles observées dans le jeu de données. On construit pour cela les diagrammes de fiabilité ci-dessous :

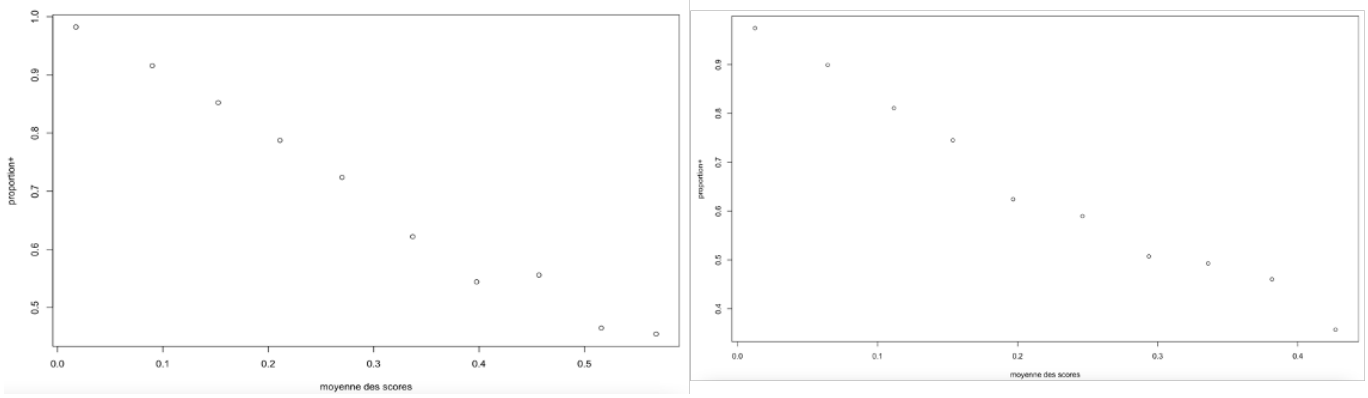


FIGURE 3.9 – Moyenne des scores en fonction du nombre de démission sur l'*échantillon d'apprentissage* (à gauche) ainsi que sur l'*échantillon test* (à droite)

Nous constatons que le modèle produit des scores bien calibrés puisque les points sont quasiment alignés sur la droite. On peut alors supposer au vu de la qualité d'ajustement de notre modèle, qu'il existe bel et bien un effet linéaire entre nos départs en démissions et les variables explicatives.

3.5.3 Mesures de la qualité prédictive du modèle

Une fois la pertinence du modèle établie, nous souhaitons dans cette section mesurer la qualité de prédiction de celui-ci. En effet, gardons en mémoire que d'une part un modèle parfaitement ajusté colle aux données d'apprentissage mais en contrepartie est faible en pouvoir prédictif sur de nouvelles données. D'autre part, l'objectif premier dans cette section est de prédire les démissions, nous souhaitons obtenir une bonne qualité prédictive de notre modèle. La mesure de performance prédictive se fait sur des données n'ayant pas servies à l'entrainement du modèle, autrement dit sur l'*échantillon test.*

Pour cela, nous confrontons les valeurs observées de notre variable réponse avec les prédictions du modèle. On construit donc une matrice de confusion qui prend la forme suivante :

Prédiction Réelle	Positive	Négative
Positive	VP	FP
Négative	FN	VN

Table 3.5 – Matrice de confusion

Où FP est le nombre de faux positifs, c'est-à-dire prédits comme positifs par le modèle mais qui ne le sont pas en réalité, VP est le nombre de vrais positifs, FN est le nombre de faux négatifs et VN est le nombre de vrais négatifs.

De cette matrice, nous tirons les différents critères de performance suivants :

— Taux d'erreur correspondant au nombre de mauvais classement,

$$\epsilon = \frac{(FN + FP)}{(VP + FN + VN + FP)}$$

— Taux de succès correspondant au nombre de bon classement ,

$$\tau = 1 - \epsilon$$

— La sensibilité correspondant à la proportion de vrais positifs est un indicateur de capacité à prédire les positifs ,

$$Sensibilité = \frac{VP}{VP + FN}$$

— La spécificité correspond au taux de vrais négatifs,

$$Spécificité = \frac{VN}{FP + VN}$$

— La précision est la proportion des vrais positifs parmi les prédictions positives, c'est la probabilité pour un individu d'être réellement positif lorsque le modèle le prédit positif,

$$Précision = \frac{VP}{VP + FP}$$

— La F-measure est la moyenne harmonique entre la sensibilité et la précision,

$$F - measure = \frac{2 \cdot Précision \cdot Sensibilité}{Précision + Sensibilité} = \frac{2 \cdot VP}{2 \cdot VP + FP + FN}$$

A noter que les critères introduits jusqu'ici sont très fortement dépendants de la règle de prévision. En effet, pour affecter la modalité à un individu l'utilisateur fixe ce que l'on appelle une règle de prévisions et notamment un seuil (généralement 0.5 dans le cas où nous considérons qu'un individu a autant de chance d'appartenir aux deux différentes modalités). Cette règle prend la forme suivante :

$$p_\beta(x) > 0,5 \text{ alors } Y = +$$

Un critère de performance présentant l'avantage d'être indépendante au seuil de prédiction est la courbe ROC. C'est un outil graphique qui met en relation le taux de vrais positifs et le taux de faux positifs en faisant varier le seuil d'affectation de la règle de prévision. De cette courbe, nous pouvons ensuite calculer l'AUC correspondant à l'aire sous la courbe. Plus l'AUC est élevé plus le modèle se rapproche du modèle parfait, correspondant à un AUC de 1.

3.5.4 Les performances de la régression logistique

Nous présentons ci-dessous les différents critères de performance de notre modèle construit à l'aide de la validation croisée :

Sensibilité	Spécificité	Précision	Taux d'erreur	AUC
0,7120	0,8032	0,9789	0,2814	0,8325

TABLE 3.6 – Critères de performance associés à la régression logistique obtenue par validation croisée sur l'*échantillon d'apprentissage et de test*

A noter que dans un souci d'affichage des meilleures performances uniquement, le seuil d'affectation a été optimisé et est de $0,074643$. Nous reviendrons dans le chapitre suivant, lors de la modélisation des promotions, sur la méthode utilisée et l'intérêt de cette approche lorsque la finalité du modèle de prédiction consiste en une affectation de classe de nos individus. En effet ici, nous nous limitons à ce que l'on appelle le *scoring* en anglais, puisque lors du modèle de calcul des engagements sociaux nous appliquons une probabilité de présence au terme à chaque individu.

3.6 Implémentation de l'agorithme CART

3.6.1 Construction de l'arbre et élaguage

Pour la construction de notre arbre de décision nous allons procéder comme explicité ci-dessous :

— Nous construisons d'abord l'arbre maximal,

— L'arbre est ensuite élagué selon des critères différents ce qui aboutit à plusieurs arbres de décision,

— Les arbres obtenus à l'étape précédente sont ensuite concurrencés et nous sélectionnons l'arbre final.

Nous commençons en construisant l'arbre de décision saturé, sans critères de nombre d'observations dans les nœuds. Celui-ci implémenté à l'aide des paramètres par défaut de R, est uniquement composé de la racine ce qui signifie que la subdivision de nos données ne réduit pas le manque d'ajustement global du modèle mesuré par l'indice de Gini d'une unité du paramètre de complexité.

Si désormais, nous gardons un nombre d'observations par nœud à 20 (paramètre par défaut de R) mais que nous abaissons le paramètre de complexité à 0 afin de construire l'arbre le plus grand possible. Nous obtenons l'arbre « surentrainé » suivant :

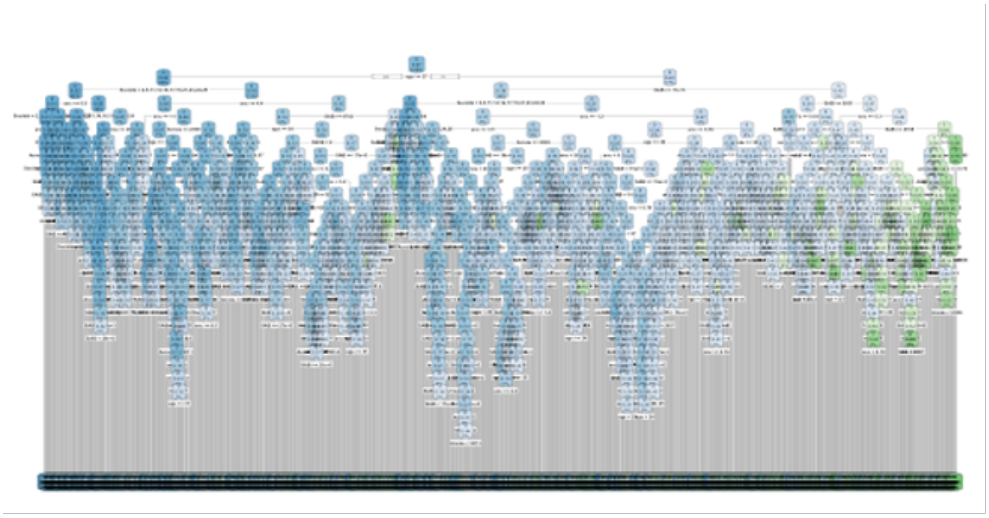


FIGURE 3.10 – Arbre de décision maximal obtenu sur l'*échantillon d'apprentissage*

On parle de « sur entrainement » lorsque qu'un arbre est beaucoup trop grand pour être lisible et trop spécifique (c'est à dire qu'il subdivise trop nos données).

Concrètement cela signifie qu'il est bien ajusté à nos données mais rappelons-le, lorsque nous avons séparé nos données en échantillons d'apprentissage et de test, il était question de compromis puisqu'un modèle dit saturé ou sur-paramétré est bien ajusté aux données observées dans l'échantillon apprentissage, réduisant ainsi l'erreur entre données prédites par le modèle et données réelles, mais va résulter d'une saturation un modèle peu robuste face à de nouvelles données telles que celles de l'échantillon test. De plus, rappelons que dans notre cas nous voulons prédire les démissions donc nous préférons créer un arbre robuste en prévision présentant des bonnes performances prédictives sur de nouvelles données.

Pour revenir à notre arbre, nous allons donc chercher le « compromis » optimal afin d'améliorer le pouvoir prédictif de notre arbre. Ce compromis se matérialise par la taille de notre arbre et donc le paramètre de complexité permettant d'arrêter l'algorithme de construction. C'est l'hyper-paramètre de l'algorithme CART, cette notion sera introduite dans la section qui suit

concernant le random forest. Pour trouver le paramètre cp optimal nous allons élaguer notre arbre selon deux possibilités de critères de sélection :

- $min\ xerror$: la valeur du paramètre de complexité minimise l'erreur de validation croisée,
- $1 - SE$: la valeur du paramètre de complexité la plus élevée dont l'erreur de validation croisée est toujours dans un SE (SE correspond au terme anglais *Standard Error* qui est une estimation de l'écart type) de l'erreur de validation croisée minimale possible.

La règle dite du « $1 - SE$ » qui nous permet de construire l'arbre ci-dessous présentant les meilleures performances :

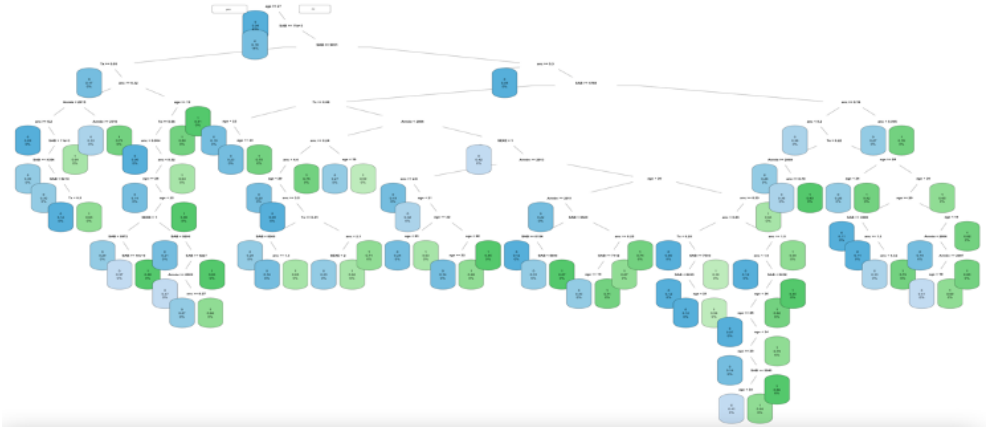


FIGURE 3.11 – Arbre de décision élagué selon la règle « $1 - SE$ »

3.6.2 Les performances de l'arbre CART

Ci-dessous, nous présentons les performances associées du modèle obtenu :

Sensibilité	Spécificité	Précision	Taux d'erreur	AUC
0,8612	0,5243	0,9588	0,1632	0,7003

TABLE 3.7 – Critères de performance associés à l'algorithme CART obtenu par validation croisée sur l'*échantillon d'apprentissage et de test*

Comme pour la régression logistique, le seuil a été optimisé à 0,1199051 afin d'afficher les critères optimaux.

3.7 Implémentation du random forest

3.7.1 Construction du modèle

Pour rappel, l'algorithme du random forest est une technique d'agrégation des arbres de décisions, en résulte donc un certain nombre d'arbres. Dans un premier temps, nous procédons au calibrage des paramètres dits hyper-paramètres (*hyper-parameters* en anglais) du

random forest. Ces hyper-paramètres ne sont pas des paramètres de modèle et, à ce titre, ils ne peuvent pas être directement entraînés à partir des données. En effet, alors que les paramètres du modèle spécifient comment transformer les données d'entrée en sortie souhaitée, les hyper-paramètres définissent comment notre modèle est structuré.

Ce procédé est appelé *hyper-parameter tuning* et permet d'améliorer les performances de notre modèle. En général, ce processus comprend les étapes suivantes :

- Définir un modèle,
- Définir la plage de valeurs possibles pour tous les hyper-paramètres (ceci peut être réalisé par l'utilisateur ou bien nous pouvons utiliser la technique du *Random Search*),
- Définir une méthode d'échantillonnage afin de déterminer les valeurs d'hyper-paramètres optimales,
- Définir un critère d'évaluation pour évaluer le modèle,
- Définir une méthode de validation croisée afin d'évaluer le modèle.

Les hyper-paramètres sont constitués pour les random forests du nombre d'arbres $ntree$ ainsi que du nombre de variables à intégrer au tirage pour chaque séparation $mtry$.

Plus $ntree$ est grand, plus la prédiction sera précise en termes d'erreur. Toutefois, $ntree$ doit être choisi suffisamment grand pour atteindre des performances satisfaisantes et suffisamment petit pour rendre les calculs réalisables (les temps de calcul de l'algorithme augmentent linéairement avec $ntree$).

Nous allons donc déterminer le nombre d'arbres adéquates, nous choisissons le nombre d'arbre optimal qui ne réduit plus le taux d'erreur de prédiction. Pour cela nous testons différentes valeurs du paramètres et réitérons plusieurs fois l'algorithme afin d'obtenir une moyenne du taux d'erreur sur l'*échantillon test*. Le graphique ci-dessous représente cette moyenne du taux d'erreur en fonction du nombre d'arbres.

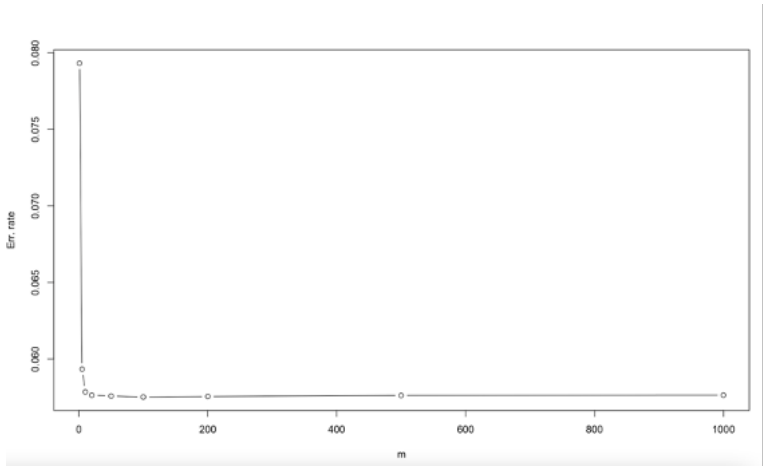


FIGURE 3.12 – Erreur de prédiction sur l'*échantillon test* en fonction du nombre d'arbres

Nous constatons qu'à partir de $ntree = 500$, les arbres additionnels n'améliorent pas les performances prédictives de la forêt aléatoire, nous limitons donc le nombre d'arbre à 500. Concernant le critère de performance AUC, au-delà de 500 arbres celui n'augmente que très légèrement ($0,8123$ pour 500 arbres et $0,8137$ pour 1500 arbres). Nous décidons, bien que le

critère AUC soit légèrement meilleur pour 1500 arbres, de ne pas augmenter le temps de calcul de notre algorithme, la plus-value étant minime.

Nous constatons par ailleurs que l'erreur OOB est inférieur pour 1500 arbre 7,56% contre 7,59% pour 500 arbres, mais là encore la plus-value est minime.

L'erreur OOB, signifie l'erreur *Out of bag*, en, effet au sein de la constrcution de chaque arbre, on ne consière que 2/3 de l'échantillon obtenu par boostrap. C'est sur le 1/3 restant que l'erreur OOB est calculée.

La validation croisée pour évaluer les performances du modèle est dans le cas du random forest pas nécessaire pour obtenir une estimation non-biaisée de celles-ci, elle nous sera utile pour l'optimisation des hyper-paramètres du modèle final.

Pour ce qui est du paramètre $mtry$, le nombre de variables à intégrer lors du tirage des variables à prendre en compte à chaque séparation, la littérature sur l'influence de ce paramètre diverge. Certains reporte que l'erreur de prédiction n'est pas affectée par différentes valeurs de $mtry$ et que les autres critères de performance telle que la sensibilité, la spécificité ou le critère AUC sont relativement stables en fonction des différentes valeurs de $mtry$. Pour d'autres, les estimations de l'importance des variables prédictives sont fortement influencées par le paramètre $mtry$. La valeur conseillée par Breiman, pour les problèmes de classification, est $sqrt(p)$.

Ainsi nous procédons à une optimisation, en testant différentes valeurs et on choisit le paramètre augmentant l'AUC un maximum.

A noter que si l'on choisit $mtry = 1$, le fractionnement est effectué aléatoirement parmi les p variables explicatives. Si au contraire, $mtry = p$, la division est réalisée dans toutes les directions possibles en incluant la totalité des variables, ce qui signifie que la construction de l'arbre est déterministe. Aussi, plus le paramètre $mtry < p$, plus l'algorithme est plus rapide.

Nous avons effectué différents tests avec un paramètre $mtry$ variant de 1 à p, il en résulte que le modèle intégrant le maximum de variables présente les meilleures performances aussi bien en termes d'erreur de prédiction que de critère AUC ($mtry = 8$ $AUC = 0,8162$ et $mtry = 2$ $AUC = 0,8039$).

A noter que ces résultats sont obtenus à l'aide d'une forêt aléatoire entrainée sur l'*échantillon d'apprentissage* et testée sur l'*échantillon test* cependant pour le choix final du modèle l'optimisation des hyper-paramètres est réitérée par validation croisée, nous obtenons les résultats suivants :

$ntree$	$mtry$
500	8

TABLE 3.8 – Hyper-paramètres du random forest obtenus par validation croisée sur l'*échantillon d'apprentissage et de test*

3.7.2 Les performances du random forest

Ci-dessous les performances associées du modèle random forest :

Sensibilité	Spécificité	Précision	Taux d'erreur	AUC
0,9952	0,0651	0,9318	0,0721	0,8156

TABLE 3.9 – Critères de performance associés au random forest obtenu par validation croisée sur l'*échantillon d'apprentissage et de test*

3.8 Implémentation du Gradient Boosting

3.8.1 Construction du modèle

Concernant le gradient boosting,les hyper-paramètres à optimiser sont les suivants :

- $nround$: correspond au nombre d'itérations à effectuer dans l'algorithme,
- max_depth : correspond à la profondeur maximale des arbres. Plus celui-ci est grand, plus le modèle est complexe mais plus le risque de sur-apprentissage est élevé,
- eta : réduction de la taille des pas pour éviter le sur-apprentissage. A chaque itération nous obtenons les pondérations des mauvais classements et ces itération sont réduites de eta,
- $gamma$: correspond à la réduction minimale de la fonction de perte requise pour effectuer une partition supplémentaire. Si la réduction ne réduit pas d'une unité $gamma$ alors le nœud n'est pas retenu,
- $colsample\_bylevel$: correpsond au ratio à appliquer aux données colonnes à intégrer dans le sous-échantillon lors de la construction de chaque arbre,
- min_child_weight : correspond à la somme minimale des poids. Si l'étape de partition de l'arbre aboutit à un nœud avec la somme des poids inférieure à min_child_weight, alors le processus de construction abandonnera le partitionnement supplémentaire,
- $subsample$: dénote la fraction d'observations à échantillonner par tiage au hasard pour chaque arbre.

Le nombre d'hyper-paramètres étant bien plus élevé que lalgorithme du *Random Forest* pour lequel nous avons testé plusieurs paramètres et comparé à chaque fois les perfomances selon l'erreur de prédiction et le critère AUC, dans ce cas précis nous décidons de nous tourner vers le package *caret* de R nous offrant la possibilité de facilement effectuer le *tuning*. Pour plus de détails, la procédure est détaillée en Annexe.1.

Le tableau ci-dessous répertorie les résultats du *tuning* obtenus par validation croisée sur l'échantillon d'apprentissage et de test :

$nrounds$	max_depth	eta	$gamma$	$colsample\_bylevel$	min_child_weight	$subsample$
1000	6	0.01	0.05	0.8	2	0.75

TABLE 3.10 – Hyper-paramètres du gradient boosting obtenus par validation croisée sur l'*échantillon d'apprentissage et de test*

3.8.2 Les performances du gradient boosting

Ci-dessous nous présentons les performances du gradient boosting évaluées par validation croisée sur l'*échantillon d'apprentissage et de test* :

Sensibilité	Spécificité	Précision	Taux d'erreur	AUC
0.8296	0.7557	0.9827	0.2389	0.8703

TABLE 3.11 – Critères de performance associés au gradient boosting obtenu par validation croisée sur l'*échantillon d'apprentissage et de test*

A noter que le seuil d'affectation a été optimisé afin de présenter les meilleures performances. Ce seuil est de 0,08046401.

3.9 L'importance des variables explicatives

Suite à la construction des différents modèles, nous pouvons mettre en évidence l'importance des variables explicatives dans chacun d'eux. Parmi les variables ayant le plus d'importance dans la prédiction des démissions, nous retrouvons, sans surprise, l'ancienneté et l'âge qui sont à chaque fois parmi les quatre variables les plus importantes. Aussi, nous retrouvons la variable le taux d'activité qui semble avoir son importance dans presque tous les modèles (sauf le gradient boosting) et dont l'importance avait déjà été évoquée dans l'analyse en composante principales.

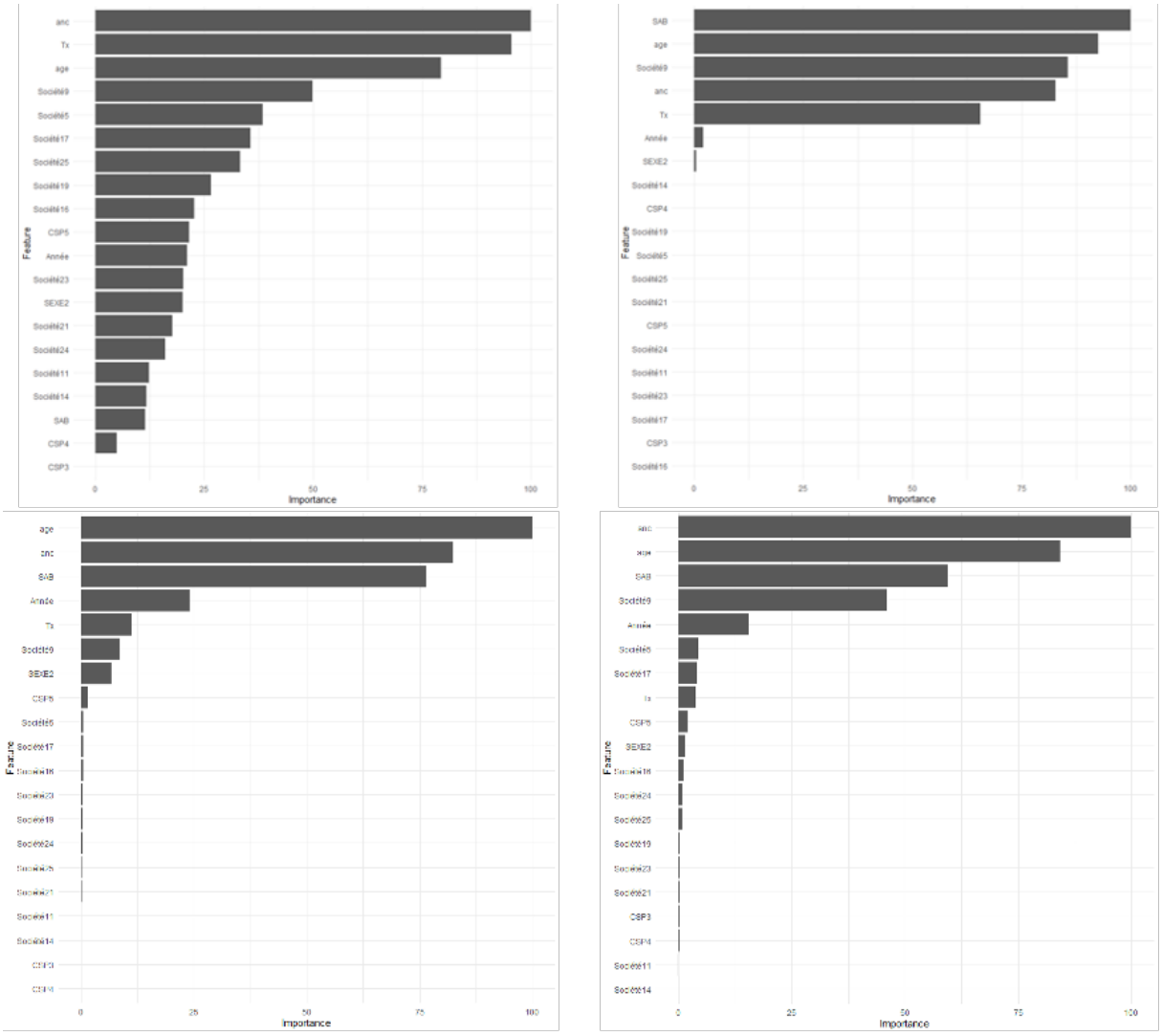


FIGURE 3.13 – Importance des variables dans la régression logistique (en haut à gauche), dans l'arbre CART (en haut à droite), le random forest (en bas à gauche) et le gradient boosting (en bas à droite)

3.10 Le choix du modèle final

Le choix du modèle se fait à l'aide de l'échantillon de validation qui n'a pas été utilisé jusqu'à présent permettant ainsi d'appréhender correctement le pouvoir prédictif de nos différents modèles. Nous présentons dans le tableau ci-dessous les différents AUC calculés à l'aide des différents algorithmes :

Algorithme	AUC
Régression logistique	0,8287
CART	0,7017
Random Forest	0,808
Gradient Boosting	**0,8306**

TABLE 3.12 – Les différents AUC calculés sur l'*échantillon de validation* selon les différents algorithmes

Nous sélectionnons le modèle maximisant l'AUC, ainsi le modèle obtenu à l'aide de l'algorithme du gradient boosting est à retenir. Nous présentons ci-dessous les courbes ROC de nos différents modèles :

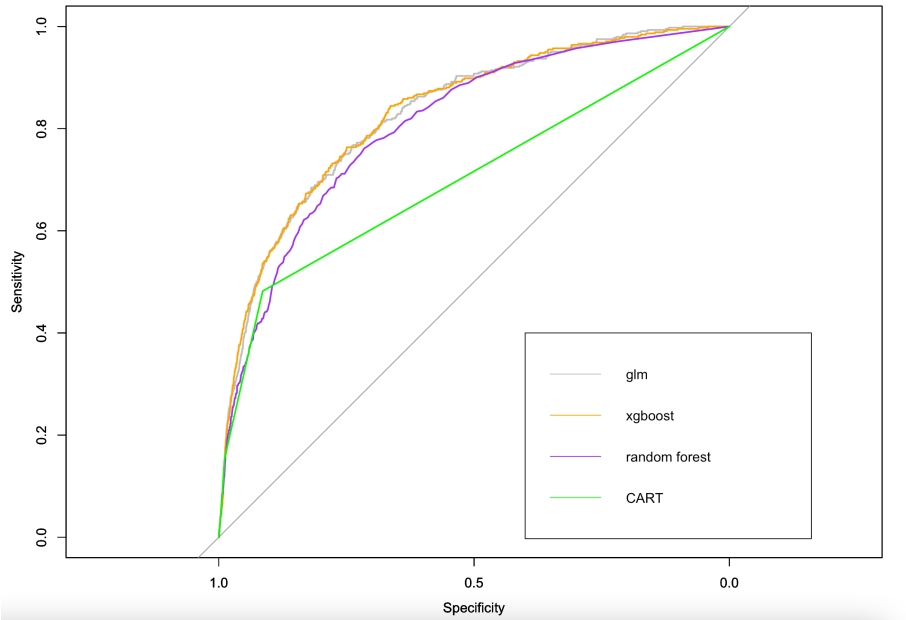


FIGURE 3.14 – Courbes ROC des différents modèles

3.11 Conclusion du chapitre

Dans ce chapitre, nous nous sommes d'abord penché sur la base servant à la modélisation du turn-over. Il s'agit de notre base globale auquelle a été ajoutée une colonne concernant la statut du salarié (autrement dit s'il démissionne au cours du prochain exercice). Une ACP réalisée sur cette base ainsi que quelques statistiques descriptives soulèvent que les démissionnaires de notre périmètres sont plutôt jeunes en âge et en ancienneté, mais révèle également l'importance du taux d'activité dans le fait de démissionner.

Nous avons, ensuite, construit différents modèles prédictifs. Selon les critères de performance adaptés aux classifications binaires et notamment le critère AUC, le modèle à retenir serait celui obtenu par l'algorithme du gradient boosting. Dans le dernier chapitre de cette

étude, nous allons procéder au calcul de l'engagement en utilisant nos modèles. Ceci nous permettra d'utiliser un indicateur plus concret (la PBO elle-même) afin de vérifier la qualité de nos modèles et surtout leur applicabilité réelle.

Chapitre 4

La modélisation de l'évolution des salaires

Dans ce chapitre nous allons, tout d'abord, nous intéresser à la modélisation du taux d'évolution des salaires. Puis, nous tacherons d'appliquer le même procédé que celui du turnover afin de modèliser les promotions cette fois-ci.

4.1 Traitement des données

Dans cette partie, nous revenons sur la manière dont est constituée la base de données servant cette fois-ci à la construction des modèles d'évolution des salaires. La base prend la forme suivante :

Année	Société	Etablissement	Matricule	Sexe	DDN	DDA	CSP	SAB $N-1$	Tx	age	anc	SAB N
2006	3	74700	4350	1	10/06/1947	04/05/1976	7	12 485.85	0.65	58.6	29.7	13 890
...	...	...	...	...	...	...	...	...	...	...	...	...

Table 4.1 – Base de données pour la modélisation de l'évolution des salaires

Là encore, nous avons besoin de deux exercices consécutifs pour reconstituer une évolution des salaires annuels bruts.

Aussi, afin de prendre en compte le taux d'augmentation des salaires uniquement lié à une évolution salariale et non pas à un changement de taux d'activité, les salaires sont donc reconstitués sur une base de temps plein. Egalement, dans un souci équivalent à celui que nous venons d'expliciter, nous exluons toutes promotions de la base de données, mises en évidence par changement de catégorie socio-professionnelle d'un exercice à l'autre.

A noter également que les salaires sont reconstitués sur la base d'une année complète, ainsi, figurent dans notre base de données un niveau de rémunération et non pas ce qu'a réellement touché le salarié dans la société durant l'année N. Par exemple, pour un salarié entrant pendant l'exercice en cours figurera dans la base le salaire annuel qu'il aurait touché s'il avait travaillé l'année compléte. Une vérification est alors apportée lors de la réception des données sur les salaires des entrants, il y a incohérence lorsque le salaire d'un salarié entrant est inférieur au SMIC annuel brut auquel cas nous demandons une reconstitution.

Aussi, pour les salariés absents pour arrêt maladie ou congés parental par exemple, ceux-ci sont intégrés à l'évaluation et les salaires sont reconstitués sur la base de ce que ces salariés devraient percevoir en cas de présence. Ceux-ci sont mis en évidence lorsque les évolutions de salaires montrent une baisse considérable, auquel cas nous demandons un salaire reconstitué.

4.2 Statistiques descriptives et analyse en composantes principales

4.2.1 Statistiques descriptives

Nous tracons le boxplot ci-dessous représentant les taux d'évolution des salaires de l'effectif stable d'un exercice à l'autre :

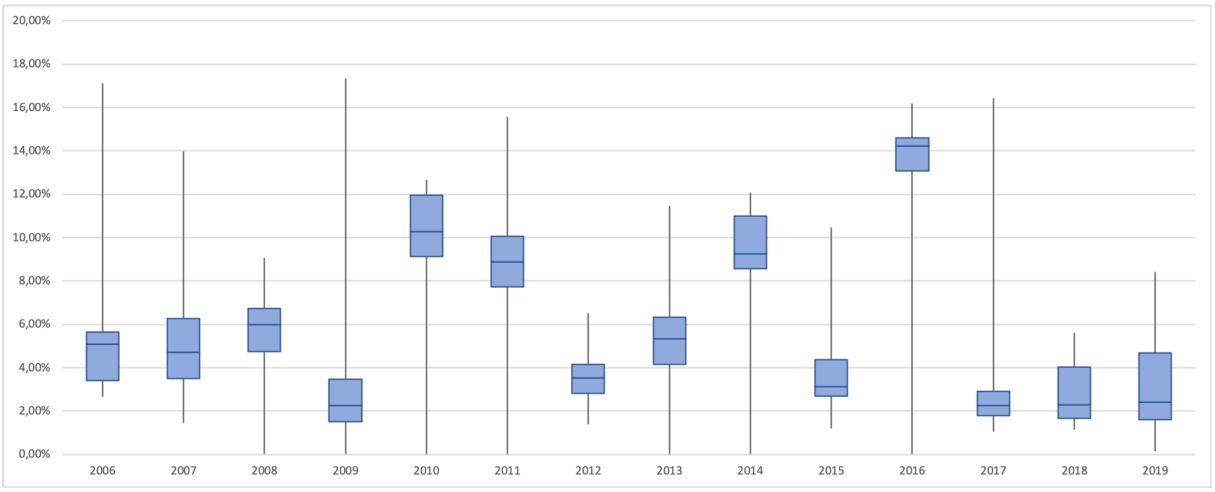


FIGURE 4.1 – Boxplot des taux d'évolution des salaires en fonction des années

On constate que les évolutions des salaires sont assez disparates en fonction des années mais également en fonction des sociétés et notamment du secteur d'activité dans lequel elles évoluent. En effet, on remarque que le secteur de la plasturgie constitue le secteur de notre périmètre présentant les évolutions salariales les plus importantes suivi par le secteur de l'indsutrie pharmaceutique puis de la métallurgie. Il semble alors important de conserver ce paramètre (société) dans la modélisation.

4.2.2 Analyse en composantes principales

Nous réalisons comme pour la partie dédiée à la modélisation du turn-over une Analyse en Composantes Principales.

Nous précisons là également que l'ACP a été centrée afin de rééquilibrer les ordres de grandeurs et faciliter l'étude des individus les uns par rapport aux autres et non par rapport à l'origine. Nous représentons ci-dessous le diagramme des valeurs propres :

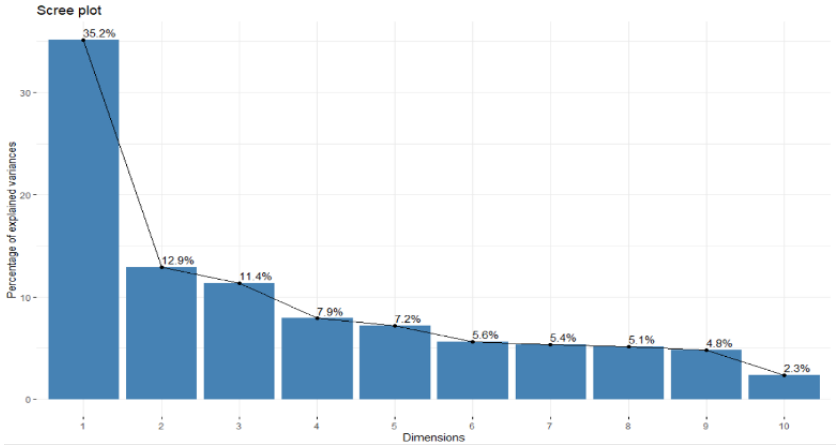


FIGURE 4.2 – Diagramme des valeurs propres

Les critères de sélection des axes ainsi que la valeur de l'inertie conservée sont synthétisés dans le tableau ci-dessous :

Critère de Kaiser	Retenir seulement les axes dont l'inertie est supérieure à l'inertie moyenne qui est égale à 1 dans le cas d'une ACP normée	Nombre d'axes retenus : 3 Valeur d'inertie conservée : 59.5%
Critère de Coude	Retenir seulement les axes avant le décrochement visible sur le diagramme	Nombre d'axes retenus : 1 Valeur d'inertie conservée : 35.2%

TABLE 4.2 – Nombres d'axes à retenir pour l'ACP selon les critères de Kaiser et de Coude

Notre choix se porte donc sur les trois premières composantes principales à retenir et nous présentons ci-dessous les contributions de nos variables selon les composantes principales.

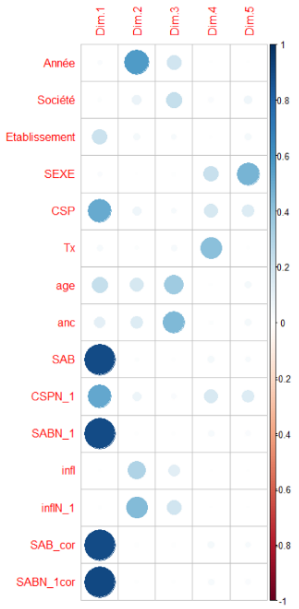


FIGURE 4.3 – Graphique de contribution des variables selon les axes de l'ACP

Le graphique ci-dessus nous permet de conclure que l'axe 1 correspond au niveau de rémunération des salariés alors que l'axe 2 s'apparente plutôt au contexte économique de l'exercice associé. On représente désormais le cercle des corrélations qui pour rappel, associe à chaque point-variable un point dont la coordonnée sur un axe factoriel est une mesure de la corrélation entre cette variable et le facteur.

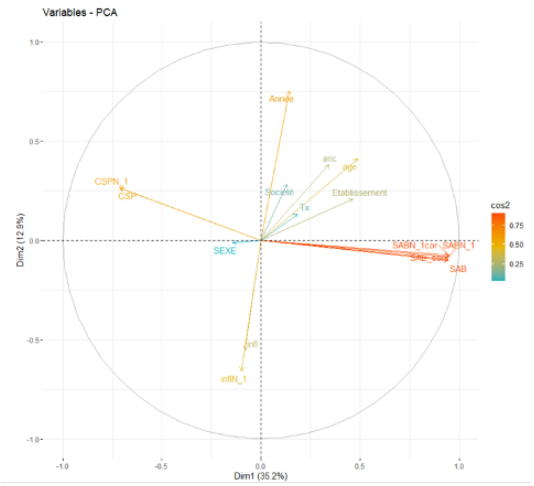


FIGURE 4.4 – Graphique de corrélation des variables sur le plan engendré par les deux premiers axes

Nous constatons bien évidemment que le salaire est corrélé avec la catégorie socio-professionnelle. Ici, nous traçons un biplot intégrant à la fois la représentation des variables mais également celle des individus :

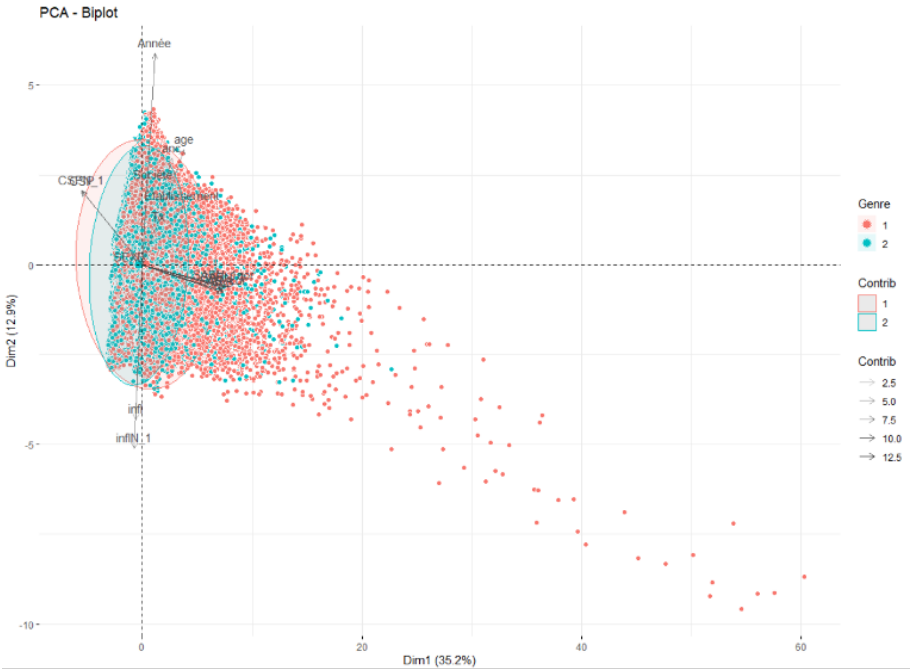


FIGURE 4.5 – Double visualisation des variables et individus sur le plan engendrépar les deux premiers axes

Cette ACP met l'accent sur le peu de corrélation entre les salaires et l'inflation. La partie qui suit a donc pour objet de vérifier ce premier postulat.

4.3 L'inflation est-elle à prendre en compte dans l'évolution des salaires ?

Comme explicité précédemment, l'hypothèse de revalorisation des salaires peut se faire soit en incluant directement l'inflation dans le pourcentage profil de carrière soit en ajoutant le taux d'inflation séparément. Quoi qu'il en soit l'inflation est, en pratique, prise en compte dans la revalorisation des salaires. Mais qu'en est-il réellement dans les faits ? La revalorisation des salaires est-elle bien indexée sur l'inflation ?

Nous allons donc réaliser, dans un premier temps, une étude de corrélation entre les augmentations de salaires observés sur l'historique de 15 années, d'une part, et l'inflation annuelle, d'autre part.

Cette étude est d'autant plus appuyée que lors de la validation des hypothèses à retenir pour l'évaluation des engagements, les différents interlocuteurs que ce soit dans les départements financiers ou dans les départements des ressources humaines des clients de Secoia, s'accordent tous pour alignés l'hypothèses d'évolution future des salaires sur l'objectif de la Banque Centrale Européenne encore fixé à 2%. Il s'agit d'ailleurs également d'une recommandation de la Banque Centrale Européenne elle-même qui recommande, pour le bon fonctionnement de l'économie, une indexation des salaires sur l'inflation.

Nous introduisons l'inflation, en temps discret comme définie dans le modèle de Wilkie introduit en 1984, qui est donnée à l'instant t par :

$$I_t = lnQ_t - lnQ_{t-1}$$

En notant Q_t l'indice des prix à l'instant t.

A noter que l'inflation annuelle est calculée comme la moyenne de l'année.

Nous récupérons les données permettant de calculer les valeurs annuelles de l'inflation en France, tirées de la série des indices de prix à la consommation ensemble des ménages en France métropolitaine base 2015 tirées de l'INSEE (série annuelle 01765178).

Nous les représentons ci-dessous :

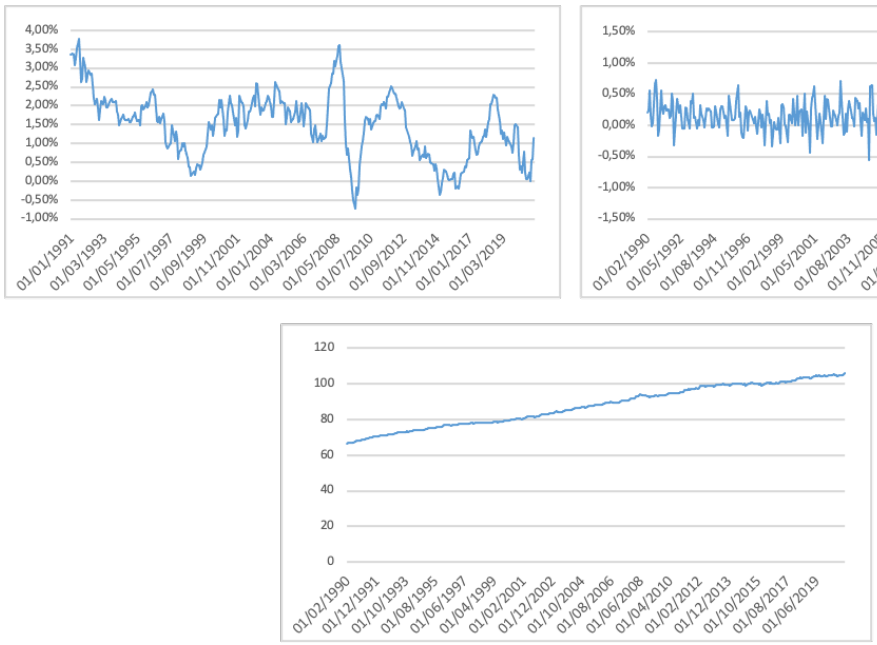


FIGURE 4.6 – Inflation annuelle (en haut à droite), inflation mensuelle (en haut à gauche) et l'indice de prix à la consommation série 01763866 tirée de l'INSEE (en bas)

Le but de l'anlyse qui suit est donc d'affirmer ou d'infirmer la liaison entre la revalorisation des salaires du périmètre étudié et l'inflation en France. Lorsque les variables sont dites liées, les variations de l'une dépendent des variations de l'autre, lorsqu'elles ne sont pas liées, elles sont indépendantes. On note dans la suite de cette section x_i l'évolution constatée d'une année à l'autre et y_i l'inflation associée aux mêmes années d'observations.

Dans la suite, nous explicitons les différents coefficients servant pour cette analyse.

4.3.1 Corrélation de Pearson

On appelle covariance la mesure de dispersion de données par rapport à leur centre (point qui a pour coordonnées les moyennes de nos deux variables). La covariance permet de quantifier la liaison entre deux variables et se calcule comme suit :

$$cov^*(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

A partir de la covariance qui très dépendante de l'unité de mesure des variables, est défini le coefficient de corrélation linéaire de Pearson. Celui-ci est égal à la covariance des séries de données centrées réduites x^* et y^* respectivement associées aux séries x et y :

$$r(x,y) = cov^*(x^*,y^*) = \frac{cov^*(x,y)}{s_x \cdot s_y}$$

Ce coefficient mesure le sens et l'intensité de la corrélation linéaire.
$|r| = 1$: il y a une relation linéaire antre les variables.
$r = 0$: il n'existe aucun lien linéaire entre les variables.

4.3.2 Corrélation de Spearman

Plutot que de mesurer une relation linéaire comme le coefficient de Pearson, le coefficient de corrélation de Spearman ne se cantonne pas à la linéarité. Il présente alors l'avantage d'être non paramétrique puisque ne repose plus sur l'hypothèse d'une distribution normale entre nos variables. Afin de le calculer, nous attribuons des rangs aux observations et on note $x_i^{'}$ le rang de la mesure x_i. Par exemple, nos observations de revalorisation des salaires prennent N valeurs distinctes, le rang de notre observation la plus faible est alors de 1 et celui de la valeur la plus haute est de N. Nous pouvons alors calculer le coefficient de corrélation de Spearman dont la formule de calcul est la suivante :

$$r_{Spearman}(x, y) = r(x^{'}, y^{'}) = 1 - \frac{6 \cdot \sum_{i=1}^{n}(x_i^{'} - y_i^{'})^2}{n(n^2 - 1)}$$

Le degré de liaison va croissant avec $|r_{Spearman}|$.

4.3.3 Présence d'ex-aequo

Dans notre cas, nous sommes en présence d'ex-aequo (l'inflation annuelle est identique pour toutes les salariés de l'exercice N). Ces ex-aequo viennent alors faussés le coefficient de Spearman. Afin de corriger cela il faut, au lieu d'attribuer des rangs à nos données ordonnées, raisonner sur la base de rang moyen. On entend par rang moyen, la moyenne des rangs pour deux valeurs identiques qui leurs auraient été attribué en cas de non ex-aequo.

En notant t_k le nombre d'apparition du rang k, le facteur de correction pour la valeur x est définit comme :

$$T_x = \sum (t_k^3 - t_k)$$

Le coefficient de corrélation de Spearman doit alors corrigé en introduisant ce facteur et devient alors :

$$r_{Spearman}(x, y) = \frac{((n^3 - n) - 6 \cdot \sum_{i=1}^{n}(x_i^{'} - y_i^{'}) - \frac{T_x + T_y}{2}}{\sqrt{(n^{-}n)^2 + (T_x + T_y) \cdot (n^3 - n) + T_x \cdot T_y}}$$

Afin d'éviter de passer par la formule corrigé du coefficient de Spearman, nous pouvons directement calculé le coefficient de Pearson sur les rangs moyens, les résultats étant équivalents.

4.3.4 Applications

Dans cette partie, nous allons tester l'indépendance de nos deux variables. Nous notons :

- H_0 : les variables sont indépendantes $r = 0$
- H_1 : les variables sont liées $r \neq 0$

Nous fixons le niveau d'acceptation d'hypothèses α à 5%. Pour déterminer le test à utiliser et le coefficient de corrélation le plus approprié à nos observations, nous considérons la loi du couple (X, Y), avec $X = évolutions\ des\ salaires$ et $Y = inflation$.

En effet, le coefficient de corrélation linéaire de Pearson s'utilise dans le cas d'une distribution bi-normale de nos variables. Nous allons donc tout d'abord vérifier cette hypothèse pour appliquer ce coefficient.

Pour cela nous appliquons le test de normalité de Shapiro & Wilk auquel nous décidons d'associer une représentation graphique des nos observations. Ce test et les résultats qui en découlent sont présents en Annexe.2. Ceux-ci indiquent que nos observations ne suivent pas une loi bi-normale.

Nous nous tournons donc vers le test non paramétrique basé sur le coefficient de corrélation corrigé (car nous sommes en présence d'ex-aequo) de Spearman définit plus haut.

Celui-ci est de $0,0595$ indiquant une faible relation entre les augmentations de salaires constatées sur notre périmètre et l'inflation.

Cependant en se penchant désormais sur le degré de significativité de la relation entre nos variables, nous calculons la loi de statistique du coefficient $r_{Spearman}$. Celle-ci peut être approximée sous H_0 sur un échantillon conséquent en terme de nombre d'observations par une loi normale. On note alors :

$$\sqrt{1-n} \cdot r_{Spearman} \sim N(0,1)$$

On en déduit alors que $\sqrt{1-n} \cdot r_{Spearman} > u_{\frac{1-\alpha}{2}}$ et on rejette donc H_0.

Ainsi le test ne nous permet pas de conclure sur une indépendance entre l'inflation et les augmentations de salaires bien que le lien paraît faible.

Dans la suite, nous allons tâcher de modéliser l'évolution futur des salaires basé sur ce que nous observons au sein de notre périmètre sans distinguer l'inflation, d'une part, et le taux de profil de carrière, d'autre part. En effet, d'après les résultats de l'analyse de corrélation ci-dessus, nous ne pouvons pas conclure à une indépendance de l'inflation et les évolutions de salaires. Cependant le lien entre ces deux variables semble somme toute assez faible. Quoi qu'il en soit, si on considère que les salaires fournis par les ressources humaines comprennent un taux d'augmentation des salaires ainsi que l'inflation nous ne modèliserons pas séparèment ces deux facteurs bien que nous aurions pu modéliser l'inflation par un modèle de Vasicek, pourtant largement répandu en terme de modélisation de taux court en temps continu.

4.4 Modélisation du taux d'augmentation des salaires

Ainsi dans le suite, nous modélisons le taux d'augmentation des salaires inflation comprise. Dans cette partie, nous allons procéder par analogie, de la même façon que pour le turn-over en utilisant d'une part des modèles linéaires généralisés et d'autre part des modèles prédictifs établis à partir d'apprentissage automatique par arbres de décisions.

Dans le cas présent, nous somme non pas face à un problème de classification mais à une régression, la variable à modéliser étant continue.

4.4.1 Modèle linéaire généralisé loi Gamma

Cette fois-ci, la fonction Gamma semble la plus appropriée afin de modéliser notre variable du salaire annuel brut de l'année suivante SAB_{N+1} pouvant prendre des valeurs continues et dont la distribution est asymétrique comme le montre le graphique ci-dessous :

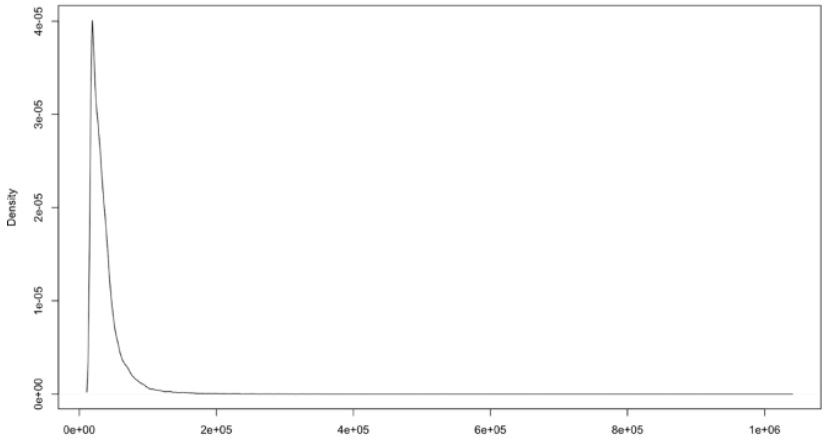


FIGURE 4.7 – Densité des salaires annuels bruts

Nous ne reviendrons pas sur les aspects théoriques derrière les modèles linéaires généralisés puisque déjà explicités dans le chapitre dédié à la modélisation du turn-over. Nous allons toutefois expliciter ce que le modèle Gamma induit et adapter les notations.

Définition : Soit Y une variable à valeurs dans $\mathbb{R}^+$ à expliquer par p variables explicatives $X = (1, X_1, \ldots, X_p)'$. Le modèle Gamma à lien logarithmique propose une modélisation de la loi $Y|X = x$ par une loi de Gamma de moyenne $\lambda_\beta(x)$ et de variance $\frac{\lambda_\beta(x)^2}{\upsilon}$ telle que :

$$log\lambda_\beta(x) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_p \cdot x_p = x^{'} \cdot \beta$$

ou encore :

$$log\lambda_\beta(x) = x^{'} \cdot \beta$$

Dans ce cas, on peut écrire :

$$\lambda_\beta(x) = \mathbb{P}_\beta(Y = 1|X = x) = exp(x^{'} \cdot \beta)$$

L'estimation des paramètres β_i du modèle se fait également par maximum de vraisemblance et la fonction de vraisemblance pour une loi log-Gamma s'écrit :

$$L(\beta, x) = \prod_{i=1}^{n} \frac{1}{\Gamma(\upsilon)} \left(\frac{\upsilon \cdot y(x_i)}{\lambda_\beta(x_i)} \right)^{\upsilon} \cdot exp\left(-\frac{\upsilon \cdot y(x_i)}{\lambda_\beta(x_i)} \right) \cdot \frac{\upsilon \cdot y(x_i)}{\lambda_\beta(x_i)}$$

4.4.2 Création du modèle et sélection des variables

Nous procédons de la même façon que pour le turn-over pour la sélection des variables et présentons ci-dessous les résultats de la construction du modèle.

Nous utilisons la fonction $drop1$ ainsi que $add1$ de R afin d'appliquer le test du rapport de vraissemblnace quant à l'ajout de nos variables dans le modèle GLM.

```
> drop1(glm(SAB ~ Année + Société + age + anc + CSP + Tx + SABN_1 + SEXE, data = Salaires[c(MYinput, MYtarget
1)], family="Gamma"(link="log")) , test =  "Chisq")
Single term deletions

Model:
SAB ~ Année + Société + age + anc + CSP + Tx + SABN_1 + SEXE
        Df Deviance     AIC scaled dev.  Pr(>Chi)
<none>       2228.9 1910623
Année    1   2356.5 1915645        5023 < 2.2e-16 ***
Société 11   3381.7 1955993       45391 < 2.2e-16 ***
age      1   2238.8 1911011         390 < 2.2e-16 ***
anc      1   2234.6 1910844         223 < 2.2e-16 ***
CSP      3   3244.1 1950593       39975 < 2.2e-16 ***
Tx       1   2392.7 1917071        6449 < 2.2e-16 ***
SABN_1   1   7151.0 2104430      193809 < 2.2e-16 ***
SEXE     1   2262.8 1911955        1333 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 4.8 – Résultat de la fonction *drop1*

Ainsi, d'après ce qui précède, toutes les variables sont significatives au seuil de 0.005 au sens du test du rapport de vraisemblance, la suppression de la moins significative « anc » n'améliore pas significativement le modèle. Nous arrêtons donc la procédure de retrait de variables ici.

Nous nous tournons ensuite vers la fonction $add1$ dont les résultats sont visibles ci-dessous :

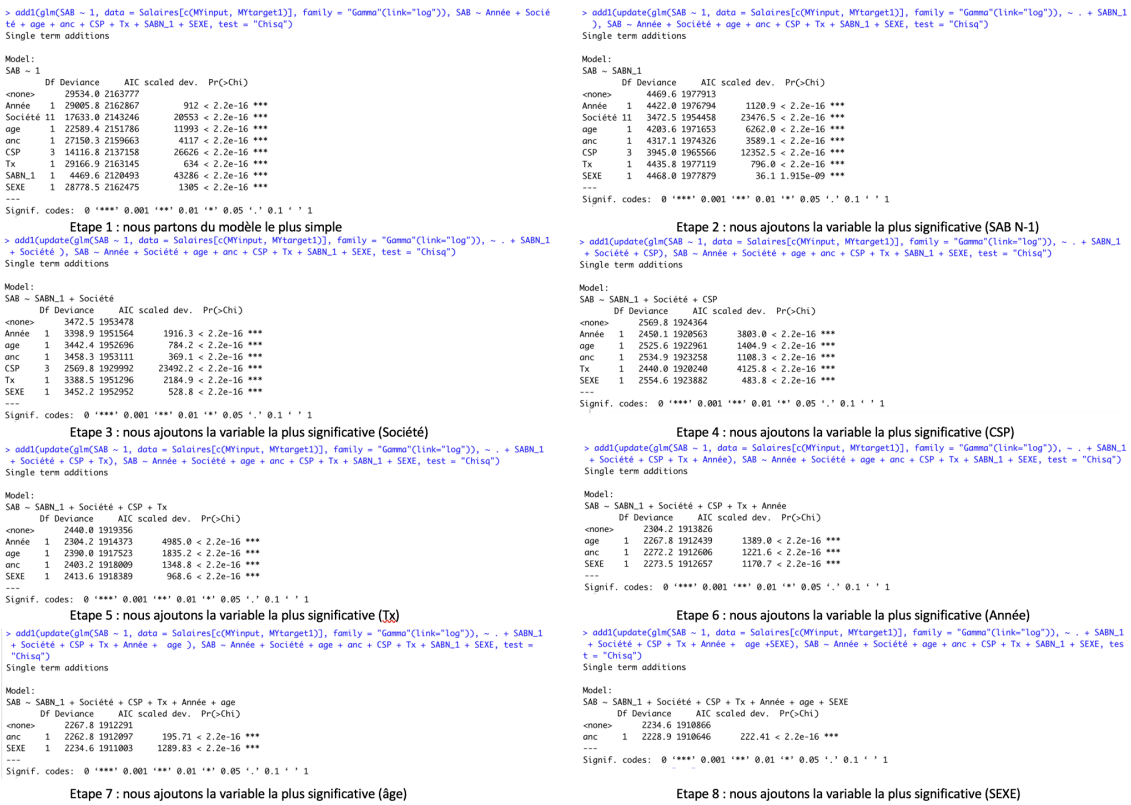


FIGURE 4.9 – Résultat de la fonction *add1*

Ainsi le modèle implémenté est synthétisé dans le tableau ci-dessous :

Variable à expliquer	Variable salaire annuel brut prenant des valeurs dans $\mathbb{R}^{+*}$
Variables explicatives	Sexe, CSP, ancienneté, âge, salaire annuel brut N-1, taux d'activité, Société, Année
Fonction lien	Fonction logarithmique
Loi	Gamma

TABLE 4.3 – Récapitulatif du modèle log-Gamma pour la modèlisation des salaires

4.4.3 Validation du modèle

Cette fois-ci, contrairement à la régression logistique, nous nous penchons sur l'analyse des résidus permettant de s'assurer de la cohérence du modèle et de l'adéquation de celui-ci en termes d'hypothèses sur le terme d'erreur.

Rappelons qu'un modèle linéaire généralisé s'écrit comme ci-dessous :

$$g(E(Y)) = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_p \cdot X_p + \epsilon_i$$

Où les résidus doivent vérifier les hypothèses suivantes :

- Indépendance : ϵ_i sont indépendants,
- Homoscédasticité : $Var(\epsilon)$ est constante.

On se doit dans ce cas d'introduire les notions de résidus de Pearson et résidus de déviance.

Les résidus de Pearson pour un individu i sont définis par :

$$r_i^P = \frac{(y_i - \hat{y}_i)}{\sqrt{Var(\hat{y}_i)}}$$

Ces résidus mesurent la contribution de chaque observation à la significativité du test découlant de la statistique du même nom, définie comme la somme des carrés des résidus de Pearson. Ainsi, si le modèle étudié est pertinent, alors la loi limite (ou exacte) de $\sum (r_i^P)^2$ est $\chi^2(n-(p+1)$.

Les résidus de la déviance pour un individu i sont donnés par : :

$$r_i^D = signe(y_i - \hat{y}_i)\sqrt{d_i}$$

Avec y_i la valeur observée de l'individu i, $\hat{y}_i$ l'estimation donnée par le modèle de l'individu i et d_i la contribution de l'individu i dans la déviance telle que $D = \sum d_i$. Ces résidus mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé.

Ces deux statistiques suivent asymptotiquement une loi de χ^2.

Nous traçons, au préalable des différents tests statistiques précités, les résidus dans le graphique qui suit :

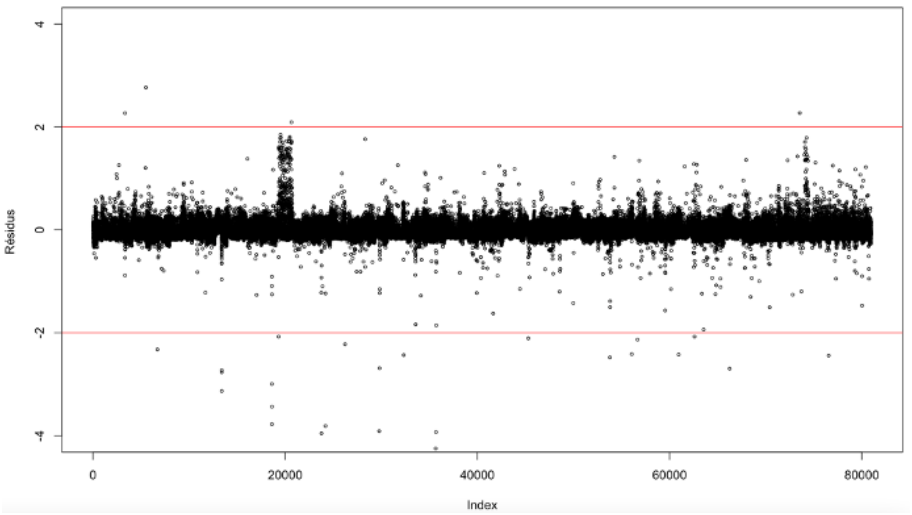


FIGURE 4.10 – Représentation graphique des résidus de Pearson du modèle

L'étude du graphique, nous permet de conclure sur l'hypothèse d'homoscédasticité des résidus qui est vérifiée, ceux-ci ne présentent globablement pas de structure/tendance indiquant une variance constante. De plus, les résidus sont majoritairement compris entre 2 et -2 hormis quelques points aberrants. On constate également que les résidus sont répartis autour de 0 et ce de manière harmonieuse (les résidus positifs semblent légèrement plus nombreux comparés aux résidus négatifs).

Ceci indique que notre modèle est satisfaisant et que toutes les hypothèses induites par les modèles linéaire généralisées semblent vérifiées. On peut alors valider ce modèle.

4.4.4 Performance prédictive du modèle

Concernant les critères de performance le critère de l'AUC ainsi que les autres critères de performance comme la sensibilité ou la spécificité ne sont plus appropriés dans le cas d'une régression. En effet, désormais, dans le cas d'une régression nous mesurons la qualité d'ajustement et le pouvoir prédictif de nos modèle par le fait que ceux-ci prédisent au « plus près des vraies valeurs ».

Ainsi, nous mesurons la racine carrée de l'erreur quadratique moyenne ($RMSE$) dans la suite de ce chapitre. Celle-ci est définie comme telle :

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

Plus la RMSE est faible, meilleur est le modèle. Nous trouvons également l'erreur absolue moyenne (MAE) définit comme la moyenne arithmétique des valeurs absolues des écarts :

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}$$

Enfin, nous pouvons reprendre une notion déjà introduite lors de la régression logistique, à savoir le coefficient de détermination que nous avions explicité en fonction de la log-vraisemblance. Celui-ci peut également s'écrire comme suit :

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

Où $\hat{y}_i$ est la valeur prédite par le modèle pour la ième observation et $\bar{y}$ la moyenne des valeurs du jeu de données.

Ci-après nous présentons les critère de performances obtenus par validation croisée et mesuré sur l'échantillon d'apprentissage de notre modèle log-gamma.

RMSE	R2	MAE
15 683 859	0.1756639	189 682.7

TABLE 4.4 – Critères de performance associés à la régression log-Gamma

4.4.5 La régression en apprentissage automatique

Pour la partie Machine Learning nous ne reviendrons pas non plus sur la théorie, nous présentons ci-dessous les résultats de nos différents modèles. L'étape dite *hyperparameters tuning* n'est pas explicitée puisque le procédé est identique à celui introduit pour le tun-over mais les hyperparamètres sont présentés en Annexe.3.

CART

RMSE	R2	MAE
16 236.96	0.6808645	10 012.83

TABLE 4.5 – Critères de performance associés à l'algorithme CART

Random Forest

RMSE	R2	MAE
4 664.885	0.9745310	1 675.494

TABLE 4.6 – Critères de performance associés à l'algorithme Random Forest

Gradient Boosting

RMSE	R2	MAE
4 273.517	0.9780396	1 619.138

TABLE 4.7 – Critères de performance associés à l'algorithme Gradient Boosting

4.5 Importance des variables dans la modélisation de l'évolution des salaires

Ci-dessous nous représentons l'importance des variables explicatives dans nos modèles :

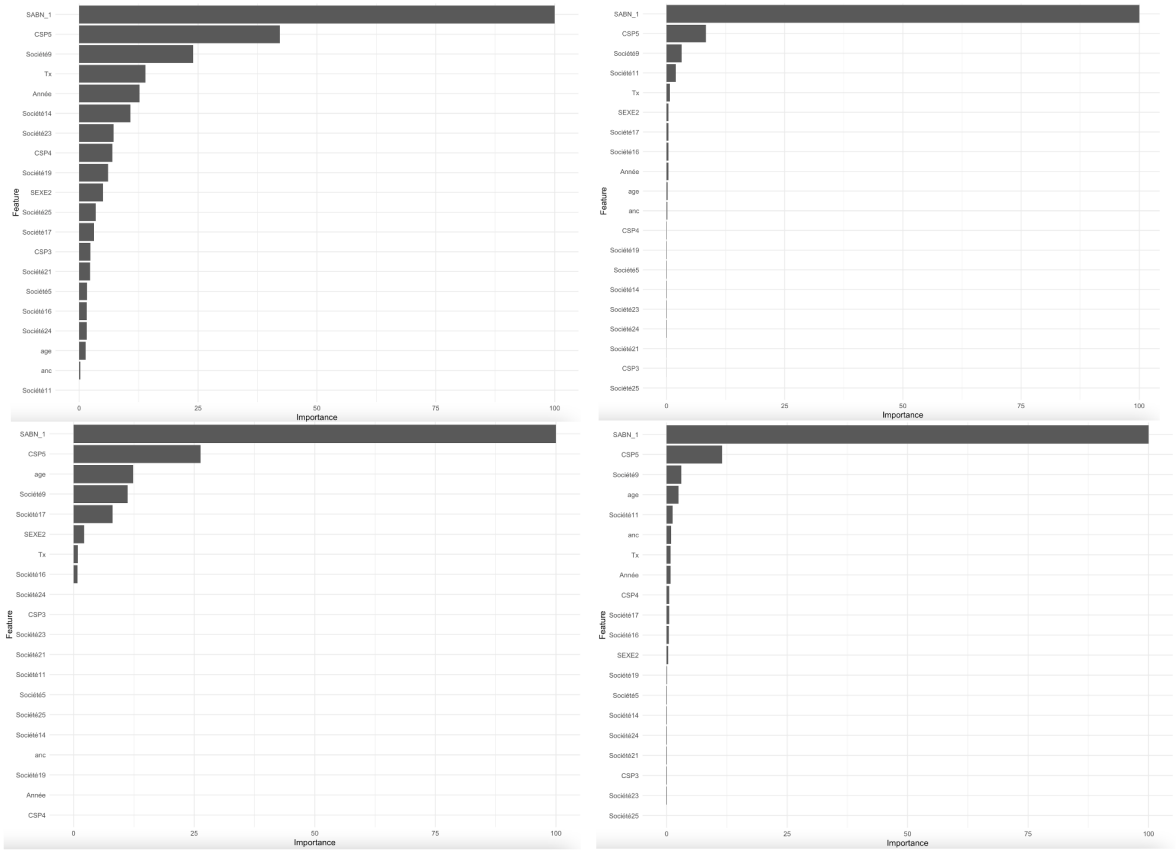


FIGURE 4.11 – Importance des variables dans le GLM (en haut à gauche), dans l'arbre CART (en haut à droite), le random forest (en bas à gauche) et le gradient boosting (en bas à droite)

Sans surprise, la variable la plus importante afin de modèliser le salaire annuel brut de l'exercice N est le salaire annuel brut de l'exercice $N-1$. La catégorie socio-professionnelle non-cadre est également importante indiquant que les non-cadres font l'objet du plus grand nombre d'augmentation. Ensuite, nous retrouvons la variable âge indiquant l'importance de différencier l'hypothèse d'évolution des salaires en fonction de l'âge.

4.6 Le choix du modèle d'évolution des salaires final

Le choix du modèle se fait à l'aide de l'échantillon de validation. Nous présentons dans le tableau ci-dessous les différents critères calculés à l'aide des différents algorithmes :

Algorithme	RMSE	R2	MAE
GLM	169 687	0.2444816	10 652
CART	17 144	0.7018842	11 413
Random Forest	**7 036**	**0.9460621**	**2 584**
Gradient Boosting	7 976	0.9321787	2 670

TABLE 4.8 – Les différents critères selon les différents algorithmes

Nous sélectionnons le modèle minimisant la RMSE ainsi que la MAE et maximisant le R2, ainsi le modèle obtenu à l'aide de l'algorithme du random forest semble être à retenir.

On note toutefois que le modèle prédictif élaboré à l'aide du GLM est, de loin, le moins performant révèlant que la loi log-gamma est finalement mal adaptée à notre problématique de prédiction de l'évolution des salaires.

4.7 Implémentation d'une probabilité d'être promu

Dans cette section, nous souhaitons modéliser les promotions afin de les prendre en compte dans le calcul des engagements sociaux.

L'historique de 15 ans nous permet de mettre au point une classification concernant cette fois-ci les promotions que nous mettons en évidence par le changement de catégorie socio-professionnelle.

Année	Société	Etablissement	Matricule	Sexe	DDN	DDA	CSP	SAB	Tx	age	anc	Promotion
2005	3	74700	4350	1	10/06/1947	04/05/1976	7	12 485.85	0.65	58.6	29.7	1
...	...	...	...	...	...	...	...	...	...	...	...	...

TABLE 4.9 – Base de données pour la modélisation des promotions

Les changements de catégorie socio-professionnelle se caractérisent généralement par une hausse des salaires annuels bruts engendrant une modification des prestations théoriques mais également un changement de barème de droits lorsque la convention collective ou l'accord d'entreprise prévoit un barème différencié selon la catégorie socio-professionnelle. Aussi les autres hypothèses peuvent différer selon la catégorie socio-professionnelle impactant à la hausse ou la baisse l'engagement. En pratique, ceci se traduit par la comptabilisation dans les comptes d'écarts actuariels d'expérience. A titre d'exemple, sur l'exercice 2019 nous estimons l'impact des promotions sur l'engagement de l'ordre de 4 870 €. Ainsi, nous souhaitons à travers cette prise en compte réduire les écarts actuarielles d'expérience et intégrer ce paramètre dans le calcul des engagements sociaux.

Bien entendu, l'impact des promotions sur la PBO dépend du nombre de promotions, nous tracons donc ci-dessous le boxplot de nos promotions en fonction des années.

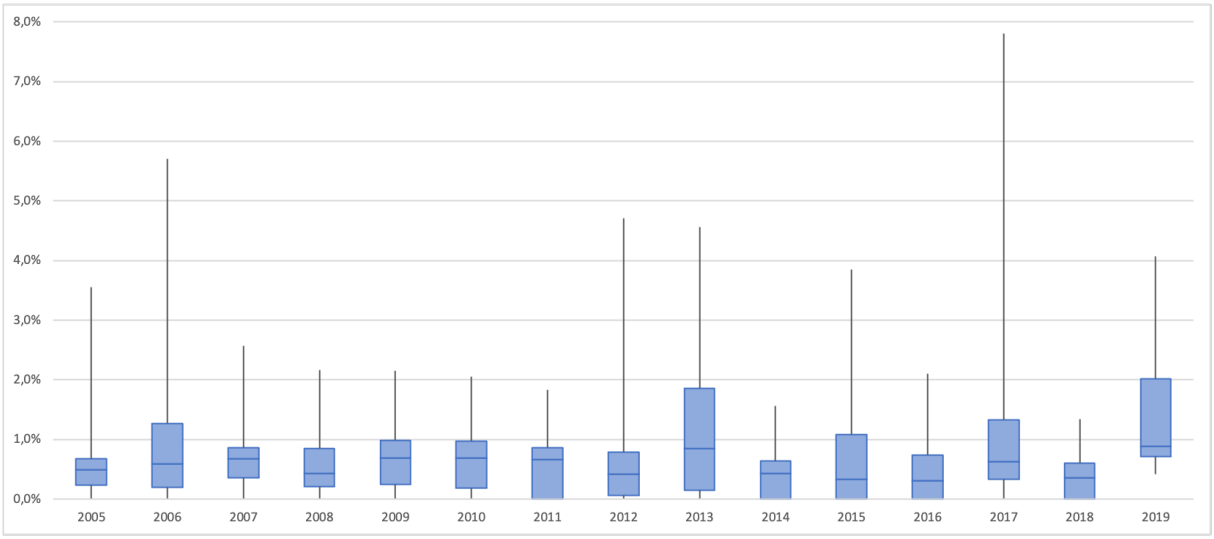


FIGURE 4.12 – Boxplot du taux de promotion en fonction des années

Comme nous pouvons le constater à l'aide du boxplot, les taux de promotion représentant le nombre de promotions par rapport à notre effectif global, sont très faibles et plutôt constants en fonction des exercices de l'historique étudié.

Comme procédé précédemment, nous réalisons désormais une ACP avant de nous lancer dans la modèlisation.

Nous représentons ci-dessous le diagramme des valeurs propres :

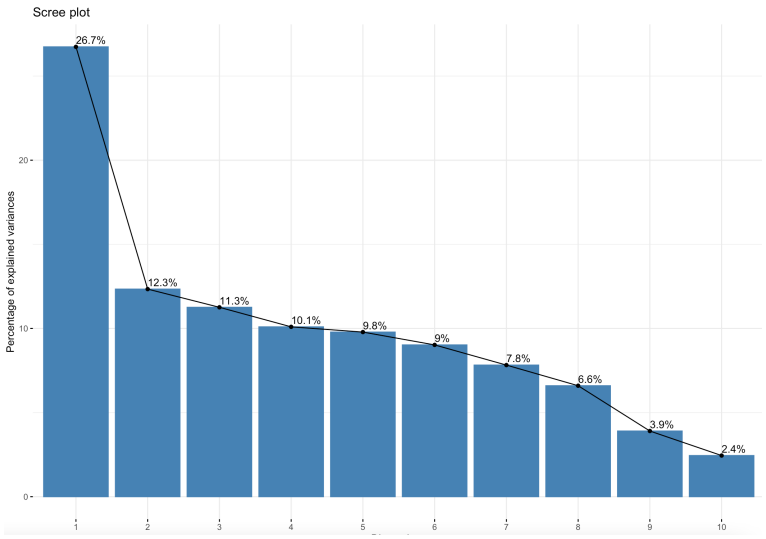


FIGURE 4.13 – Diagramme des valeurs propres

La lecture graphique nous permet de remplir la table suivante :

Critère de Kaiser	Retenir seulement les axes dont l'inertie est supérieure à l'inertie moyenne qui est égale à 1 dans le cas d'une ACP normée	Nombre d'axes retenus : 4 Valeur d'inertie conservée : 60.4%
Critère de Coude	Retenir seulement les axes avant le décrochement visible sur le diagramme	Nombre d'axes retenus : 1 Valeur d'inertie conservée : 26.7%

TABLE 4.10 – Nombres d'axes à retenir pour l'ACP selon les critères de Kaiser et de Coude

Nous conservons donc 4 axes. Ce choix est d'autant plus appuyé que la variable « promotion » qui nous intéresse est bien représentée sur la composante principale 4.

Nous traçons également le graphique de corrélation des variables ci-dessous permettant de mettre en évidence le peu de corrélation entre la variable promotion et les autres variables :

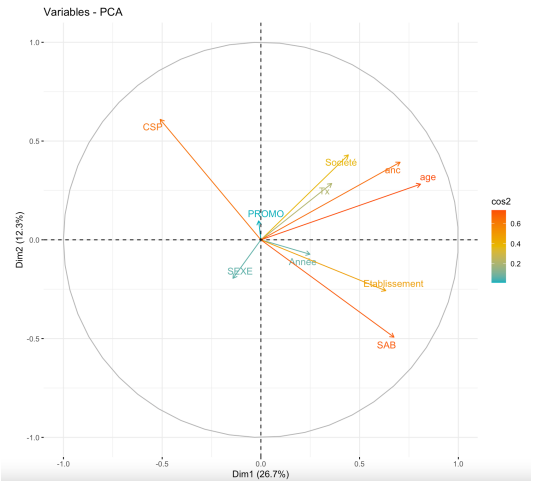


FIGURE 4.14 – Graphique de corrélation des variables sur le plan engendré par les deux premiers axes

Ici, nous traçons un biplot intégrant à la fois la représentation des variables mais également celle des individus :

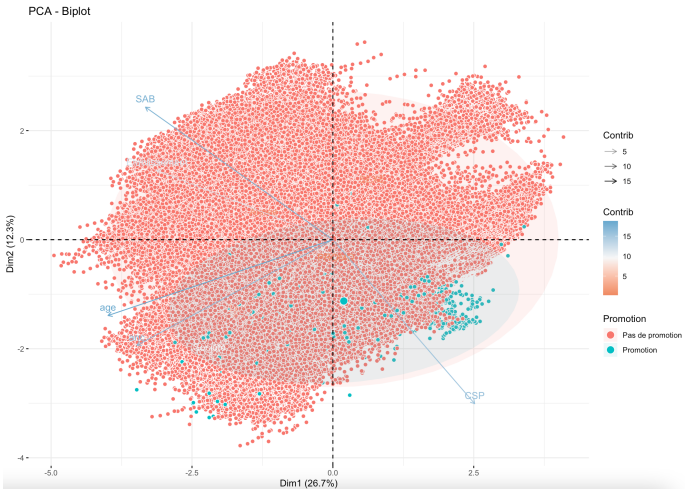


FIGURE 4.15 – Double visualisation des variables et individus sur le plan engendré par les deux premiers axes

Dans la suite, nous explicitons l'implémentation des différents modèles. Nous ne reviendrons pas sur les méthodes et algorithmes utilisés puisque sont identiques à celles et ceux utilisés précédemment. Nous sommes comme pour le turn-over face à une classification, à la seule différence que cette fois-ci, nous ne conservons plus les probabilités d'affectation aux différentes classes (bien que toujours utiles) mais nous conservons la classe des individus elle-même.

Comme évoqué, les taux de promotion sont faibles, nous nous devons alors d'introduire dans la partie qui suit la notion de déséquilibre.

4.7.1 Le déséquilibre

L'un des problèmes les plus difficiles lors de l'établissement d'un modèle prédictif se produit lorsque les classes de notre variable à expliquer présentent un grave déséquilibre. Ce problème est malheureusement souvent la norme.

Une conséquence de ceci est que les performances prédictives sont généralement très biaisées par rapport à la classe comportant les plus petites fréquences. Par exemple, si les données ont une majorité de données appartenant à la première classe et très peu dans la deuxième classe, la plupart des modèles prédictifs maximiseront la précision en prédisant uniquement la première classe. En conséquence, il y a généralement une grande sensibilité mais une faible spécificité. De plus, en présence de fort déséquilibre entre les classes les éléments de la diagonale de la matrice de variance-covariance de l'estimateur maximum de vraisemblance vont se rapprocher de 0 ce qui conduit à une augmentation de la variance des estimateurs.

C'est malheureusement ici notre cas, ci-dessous nous présentons nos effectifs en fonction de nos deux différentes modalités par année d'exercice afin de visualiser ce déséquilibre :

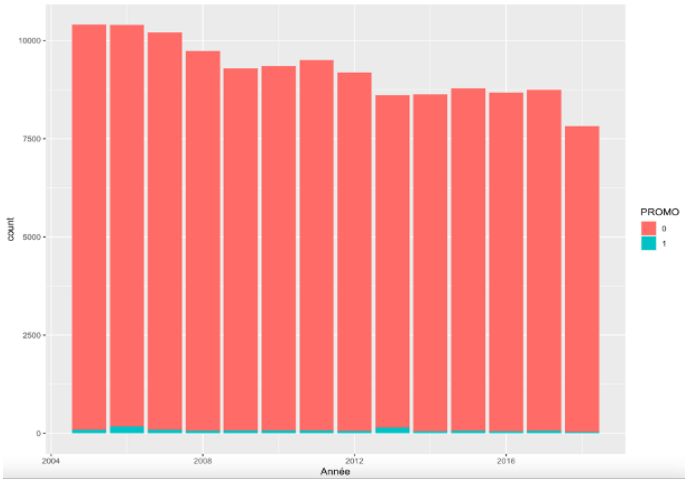


FIGURE 4.16 – Histogramme de notre effectif en fonction des années

L'effectif en rouge représente les salariés non promus lors de l'exercice en question, celui en bleu correspond aux promotions. Le déséquilibre est alors flagrant et le rapport de disproportion est de 0.008887883 pour notre base totale.

Au premier abord, nous constatons que les prévisions associées à nos différents modèles retenus modélisent d'ailleurs uniquement la classe majoritaire (avec un seuil par défaut à 0.5) et néglige la classe minoritaire, conséquence même de cette disproportion. Il est donc impératif de corriger ce déséquilibre afin d'apprécier la qualité prédictive de nos différents modèles. Se présentent alors à nous deux possibilités :

- au **niveau des données** : nous pouvons effectuer un ré-échantillonnage afin de rééquilibrer nos données,
- au **niveau algorithmique** : nous pouvons modifier les règles de prévision afin de tenir compte du déséquilibre.

4.7.2 Correction algorithmique

Dans un premier temps, nous allons optimiser le seuil d'affectation de notre modèle. Celui-ci est défini à 0,5 par défaut lorsque l'on traite d'une classification binaire, on suppose qu'un individu a autant de chance d'appartenir aux deux différentes classes. Si $p_\beta(x)$ est la probabilité d'être positif fournie par le modèle, la règle d'affectation usuelle s'écrit :

$$\text{Si } p_\beta(x) > \theta \text{ Alors } Y = + \text{ Sinon } Y = -$$

On cherche donc à optimiser cette règle en jouant sur le paramètre θ de façon à améliorer le pouvoir prédictif de notre modèle.

Pour cela, nous pouvons opter pour l'optimisation du seuil à l'aide de la courbe ROC. En effet, le seuil est choisi comme un compromis entre la sensibilité (taux de vrais positifs) et la spécificité (taux de vrais négatifs). Graphiquement, sur la courbe ROC, ce seuil correspond au point se situant le plus proche du coin en haut à gauche du graphique.

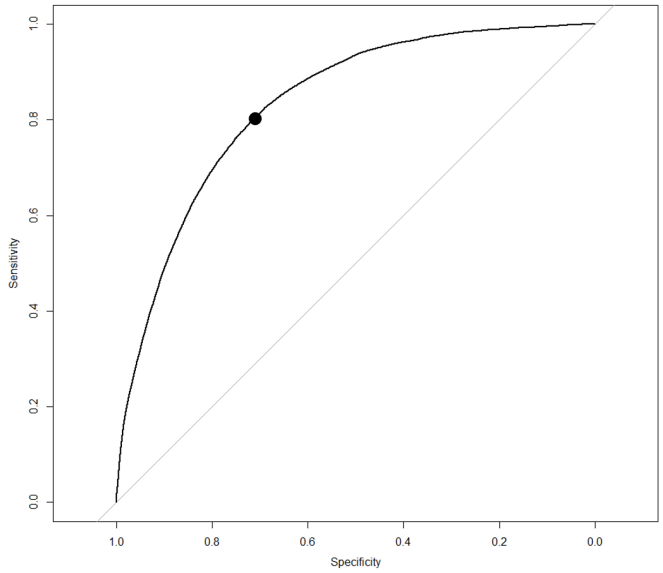


FIGURE 4.17 – Courbe ROC et seuil optimal

Ce choix ne s'avère pas toujours le meilleur selon la problématique à laquelle nous sommes confronté. Il s'agit donc de jouer sur le seuil d'affectation afin d'obtenir les performances selon les métriques souhaitées. En effet, lorsque des disparités importantes de coût existent entre les faux négatifs et les faux positifs, il peut être essentiel de minimiser l'un des types d'erreur de classification. Par exemple, dans un contexte de détection du turn-over pour reprendre notre problématique du Chapitre 3, si nous souhaitions convertir les probabilités en valeur binaire, on préférera minimiser en priorité les faux positifs puisque reviendrait à maximiser le turn-over et donc à réduire injustement et de manière considérable le montant de l'engagement.

4.7.3 Correction au niveau des données

L'autre option dont nous disposons afin de corriger le déséquilibre est le ré-échantillonnage. Le ré-échantillonnage consiste en un redressement de l'*échantillon d'entrainement* de façon à augmenter la fréquence de la classe minoritaire ou, à contrario, de diminuer celle des classes majoritaires. La redéfinition de la distribution des données est dans ce cas primordiale et représente un moyen de contourner les erreurs de prédictions engendrées par le déséquilibre. Parmi les différentes techniques d'échantillonnage existantes on trouve : le sous-échantillonnage, en anglais *under-sampling* et le sur-échantillonnage, *over-sampling*.

Il existe plusieurs techniques de sur-échantillonnage tel que le sur-échantillonnage aléatoire qui copie des données de la classe minoritaire mais il existe également des tehchniques basée sur la notion de voisinage telle que SMOTE par exemple. SMOTE signifie *Synthetic Minority Oversampling Technique*. Il crée de nouveaux échantillons synthétiques en utilisant un algorithme k-plus proche voisin pour équilibrer l'ensemble de données.

On trouve aussi des méthodes basées sur la combinaison du sur-échantillonnage et du sous-échantillonnage, toutefois nous nous limitons uniquement au sur-échantillonnage ainsi qu'au sous-échantilonnage dans ce mémoire. De plus, nous pouvons lors de l'implémentation des

différents modèles jouer sur le ratio désiré entre nos différentes classes. Afin de simplifier les procédés nous nous limitons à un ratio à 50% dans la suite.

A noter que ceci n'a pas été fait dans le chapitre lié à la modélisation du turnover puisque, comme nous l'avons explicité, nous ne nous intéressons pas à la même chose en termes d'*output* du modèle. En effet, le déséquilibre étant intrinsèque, lorsque l'on s'intéresse à la probabilité d'appartenance à une classe le fait de ré-échantillonner l'échantillon d'apprentissage dénature ces probabilités (soit par l'ajout d'observations de la classe minoritaire *over-sampling*, soit par le retrait d'observations de la classe majoritaire *under-sampling*).

Aussi, il est important de le souligner puisque nos modèles sont entrainés par validation croisée, le ré-échantillonnage doit être fait sur l'*échantillon d'apprentissage* uniquement et non sur l'*échantillon test*. Ceci est à prendre en compte dans la validation croisée, la fonction $train$ du package $caret$ nous permet d'effectuer le sur-échantillonnage ou sous-échantillonnage à l'intérieur de la validation croisée et non avant.

4.8 Application de la correction du déséquilibre

4.8.1 Régression logistique

Le modèle retenu, obtenu par validation croisée est synthétisé dans le tableau qui suit :

Variable à expliquer	Variable binaire indiquant le fait d'être promu ou non (pour rappel 0 non promu et 1 pour promotion)
Variables explicatives	CSP, ancienneté, âge, salaire annuel brut, taux d'activité, Société, Année
Fonction lien	Fonction logistique
Loi	Binomiale

TABLE 4.11 – Récapitulatif du modèle de régression logistique pour la modèlisation des promotions

La variable Sexe n'est pas prise en compte dans le modèle suite à la procédure de sélection des variables dont les résultats figurent en Annexe.4.

Nous présentons ensuite, ci-dessous, les odds ratio du modèle ainsi obtenu dans le tableau suivant :

Variables	OR	95% CI	p-value
Anc	1.02	1.01, 1.03	<0.001
Année	1.13	1.11, 1.15	<0.001
Age	1.04	1.03, 1.05	<0.001
Tx	0.08	0.02, 0.24	<0.001
Société			
2	—	—	
5	1.64	1.07, 2.53	0.024
9	0.22	0.15, 0.31	<0.001
11	0.00	0.00, 0.00	>0.9
14	0.66	0.30, 1.74	0.3
16	2.30	1.63, 3.20	<0.001
17	0.14	0.10, 0.19	<0.001
19	0.25	0.16, 0.37	<0.001
21	2.55	1.38, 5.07	0.004
23	0.56	0.33, 1.01	0.045
24	1.19	0.81, 1.77	0.4
25	0.61	0.36, 1.11	0.090
SAB	1.00	1.00, 1.00	<0.001

TABLE 4.12 – Odds Ratio et Intervalles de confiance associés à la régression logistique entrainée sur l'*échantillon d'apprentissage et de test*

Concernant la pertinence du modèle comme pour la modélisation du turn-over, nous calculons d'une part le R^2 de McFadden égale à 19% et d'autre part le test de rapport de vraissemblance indiquant la significativité au seuil de 0.005 de notre modèle.

Nous présentons dans le tableau ci-dessous les critères de performances obtenus selon la correction du déséquilibre :

Correction du déséquilibre	**Sensibilité**	**Spécificité**	**Précision**	**Taux d'erreur**	**AUC**
Seuil d'affectation	0,78281	0,84021	0,99819	0,2167	0,8893
Sous-échantillonnage	0,79804	0,87533	0,99861	0,2013	0,9037
Sur-échantillonnage	0,79824	0,88586	0,99873	0,201	0,9048

TABLE 4.13 – Critères de performance des régressions logistiques selon l'algorithme de correction du déséquilibre

4.8.2 Arbre de décision

Comme pour la partie GLM, nous allons effectuer un ré-échantillonnage afin de palier au profond déséquilibre et nous allons entrainer les différents arbres obtenus par validation croisée afin d'éliminer le biais induit par l'échantillonnage. A noter que pour chaque modèles le paramètre de complexité a été optimisé par la méthode « $1 - SE$ » qui donne à chaque fois les meilleures performances.

Les performances ainsi obtenues sont présentées ci-dessous :

Correction du déséquilibre	Sensibilité	Spécificité	Précision	Taux d'erreur	AUC
Seuil d'affectation	0,37401	0,94346	0,99414	0,0616	0,7234
Sous-échantillonnage	0,80472	0,89552	0,99885	0,1945	0,8856
Sur-échantillonnage	0,82006	0,92274	0,99916	0,179	0,8849

TABLE 4.14 – Critères de performance des algorithmes CART selon l'algorithme de correction du déséquilibre

4.8.3 Random Forest

Nous présentons les critères de performances des différents random forest obtenus par validation croisée ci-dessous. A noter que les hyper-paramètres de ces modèles figurent en Annexe.5.

Correction du déséquilibre	Sensibilité	Spécificité	Précision	Taux d'erreur	AUC
Seuil d'affectation	1	0,99972	0,96936	0,0003	0,8942
Sous-échantillonnage	0,99997	0,04556	0,81430	0,1841	0,9271
Sur-échantillonnage	1	0,24967	0,97329	0,0265	0,9230

TABLE 4.15 – Critères de performance des algorithmes random forest selon l'algorithme de correction du déséquilibre

4.8.4 Gradient Boosting

Nous présentons dans le tableau ci-dessous les performances évaluées par validation croisée des différents algorithmes de gradient boosting implémentés. Les hyper-paramètres de ces modèles figurent également en Annexe.5.

Correction du déséquilibre	Sensibilité	Spécificité	Précision	Taux d'erreur	AUC
Seuil d'affectation	0,9344	0,4413	0,9909	0,0740	0,9335
Sous-échantillonnage	0,81408	0,94381	0,99939	0,1848	0,9269
Sur-échantillonnage	0,85368	0,97015	0,99969	0,1453	0,9302

TABLE 4.16 – Critères de performance des algorithmes gradient boosting selon l'algorithme de correction du déséquilibre

4.9 Importance des variables dans la modélisation des promotions

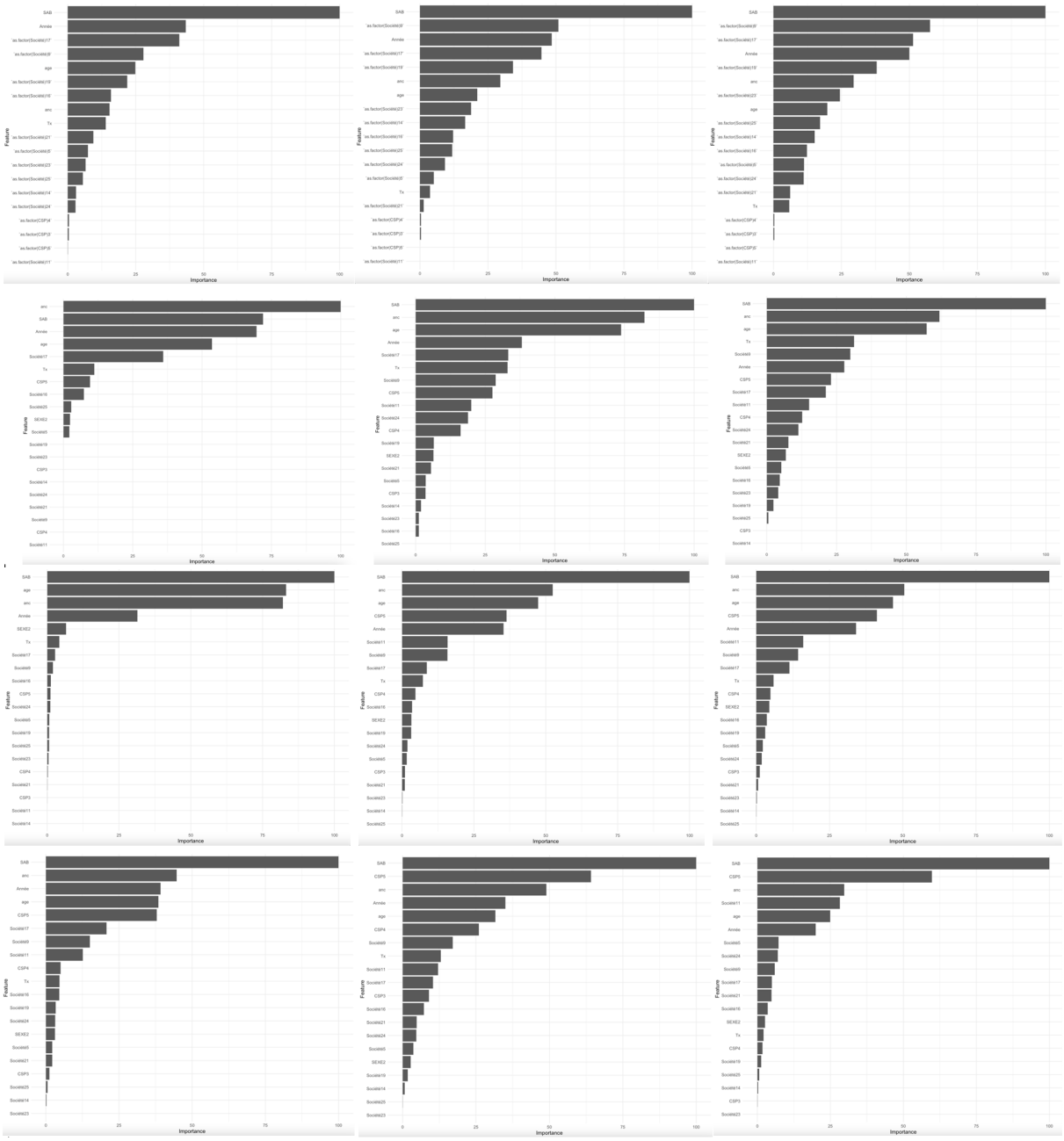


FIGURE 4.18 – Importances des variables de nos différents modèles de promotion (régréssion logistique : première ligne), CART : deuxième ligne, Random Forest : troisième ligne, Gradient Boosting : dernière ligne)

On constate que le salaire annuel brut est globalement la variable la plus importante de nos différents modèles suivie de l'ancienneté ainsi que de l'âge.

4.10 Le choix final du modèle promotion

Le choix du modèle se fait également à l'aide de l'échantillon de validation. Nous présentons dans le tableau ci-dessous les différents AUC calculés à l'aide des différents algorithmes[1] :

Algorithme	AUC
Régression logistique - optimisation du seuil	**0,8969**
Régression logistique - sous-échantillonnage	0,8908
Régression logistique - sur-échantillonnage	0,8948
CART - optimisation du seuil	0,7013
CART - sous-échantillonnage	0,8636
CART - sur-échantillonnage	0,8194
Random Forest – optimisation du seuil	0,8442
Random Forest – sous-échantillonnage	0,8767
Random Forest – sur-échantillonnage	0,8452
Gradient Boosting – optimisation du seuil	0,8876
Gradient Boosting – sous-échantillonnage	0,8637
Gradient Boosting – sur-échantillonnage	0,8776

TABLE 4.17 – Les différents AUC selon les différents algorithmes et la méthode de correction du déséquilibre utilisée

Nous sélectionnons le modèle maximisant l'AUC, ainsi le modèle obtenu à l'aide de la régression logistique semble être le modèle à retenir. Nous présentons ci-dessous les courbes ROC de nos différents modèles :

1. le critère AUC est indépendant du seuil, nous spécifions uniquement « optimisation du seuil » afin de différencier la méthode de correction du déséquilibre de nos données d'apprentissage

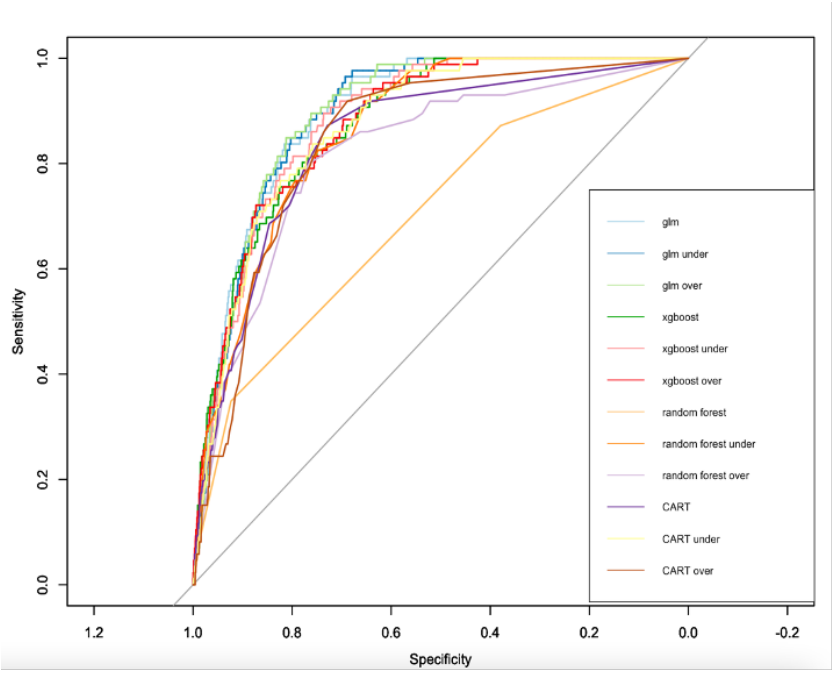


FIGURE 4.19 – Courbes ROC des différents modèles

4.11 Conclusion du chapitre

Dans ce chapitre, nous avons construit différents modèles prédictifs pour d'une part l'évolution des salaires et d'autre part les promotions.

Selon les critères de performance adaptés cette fois-ci aux régressions, le modèle à retenir pour prédire l'évolution des salaires serait obtenu par l'agorithme du random forest. Pour ce qui est de la prise en compte des promotions, le modèle obtenu par régression logistique à l'aide du changement de seuil d'affectation semble être le modèle à retenir. Toutefois, comme pour le turn-over nous utiliserons la PBO comme indicateur supplémentaire dans le chapitre suivant.

Chapitre 5

Prise en compte dans le calcul de l'engagement

Dans un premier temps, nous allons détailler les procédés mis en place afin de prendre en compte les différents modèles prédictifs construits jusqu'à présent. Puis, nous allons projeter les différents éléments de nos modèles prédictifs et nous allons confronter leurs impacts sur le montant de l'engagement pour conclure quant à cette étude.

5.1 Le modèle « PUC service prorata »

Afin de prendre en compte, les trois différents modèles construits jusqu'à présent, il faut adapter le modèle « PUC service prorata » ayant servi jusqu'à présent.

En *input* du modèle, nous retrouvons le fichier de la démographie (les salariés présents à la date de clôture) ainsi que les différentes hypothèses nécessaires au calcul des engagements sociaux tels que les barèmes de droits, les tables de survie, etc...

A la différence de notre modèle « PUC service prorata » servant de base au calcul des passifs sociaux, il faut, pour la prise en compte des modèles de turn-over, d'évolutions des salaires ainsi que des promotions, projeter et surtout garder en mémoire les différents paramètres intervenant dans le calcul tels que l'âge, l'ancienneté ou encore la durée résiduelle à chaque fin d'exercice et ce jusqu'au départ en retraite des salariés. Ceci se fait à l'aide de différentes matrices comprennant un nombre de ligne égale au nombre de salariés à la date de clôture et un nombre de colonne égale au maximum du nombre d'années résiduelles avant l'obtention du taux plein et du départ en retraite.

Le schéma ci-dessous synthétise le fonctionnement du modèle de calcul des passifs sociaux et permet de mettre en avant l'ajout des différents modèles prédictifs et l'endroit où ils interviennent :

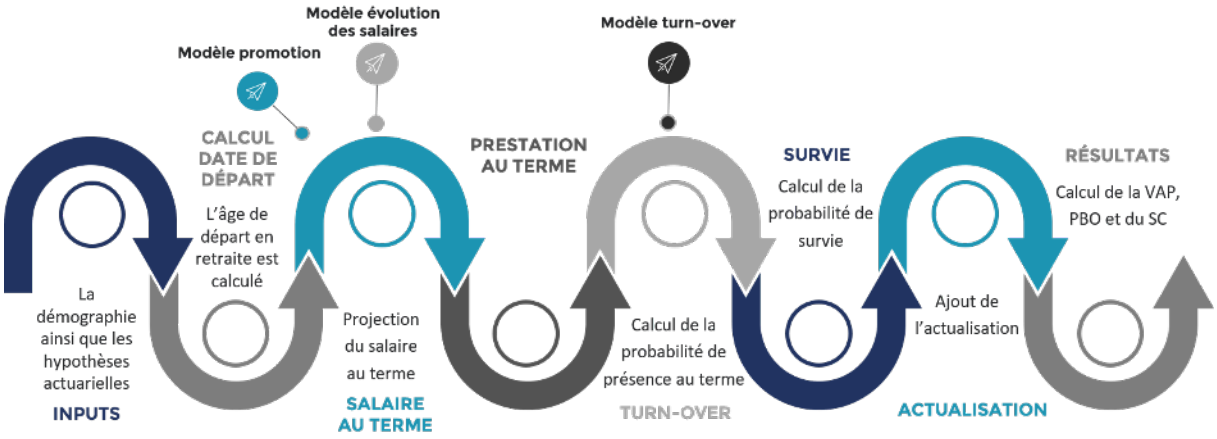


FIGURE 5.1 – Fonctionnement de l'outil de calcul des IFC

Il est à noter que le modèle qui prédit les promotions intervient avant celui des salaires afin de prendre en compte celles-ci dans l'évolution des salaires. En effet, dans notre cas, les hypothèses usuelles d'évolutions des salaires ne différent pas selon les catégories socio-professionnelles mais cela peut tout à fait être le cas. Nous laissons ainsi la possibilité d'utiliser simultanément les trois modèles prédictifs différents dans le calcul des engagements sociaux.

5.2 Les projections de la probabilité de turn-over

Nous représentons les probabilités de présence correspondant à l'inverse de la probabilité de turn-over dans le temps (jusqu'au départ en retraite du dernier salarié) et par individu. Ces projections permettent de nous faire une idée des potentiels impacts de nos modèles sur l'engagement final.

Rappelons-le, nous avons vu dans la section dédiées aux sensibilités que l'effet du turn-over dépend de la durée résiduelle du salarié. Nous obtenons donc les probabilités projetées suivantes :

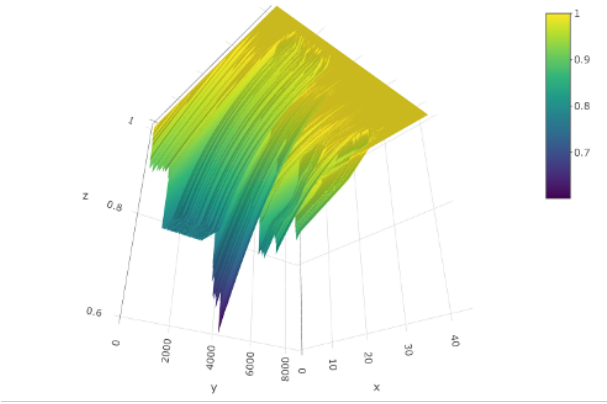


FIGURE 5.2 – Projections des probabilités du turn-over à l'aide des tables de turn-over classiques

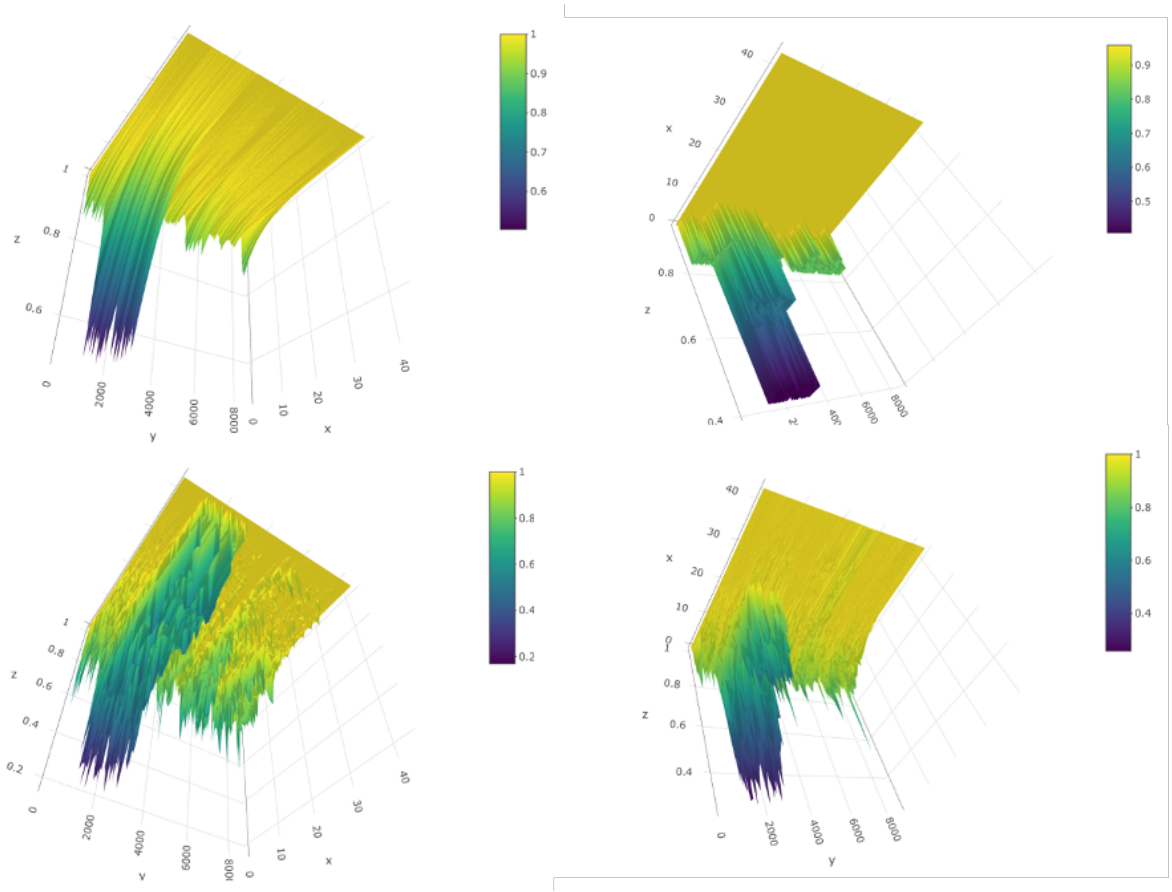


FIGURE 5.3 – Projections des probabilités du turn-over à l'aide des modèles prédictifs (régression logistique : en haut à gauche, arbre : en haut à droite, random forest : en bas à gauche et gradient boosting : en bas à droite)

Nous rappelons également que les tables de turn-over servant aux calculs de base sont des tables décroissantes par âge s'annulant soit à partir de 60 ans soit à partir de 55 ans (en fonction des sociétés). Elles ont été calibrées sur les observations empirique des 15 derniers exercices afin d'être en cohérence avec les modèles construits sur ce même historique et figurent en Annexe.7.

5.3 Les résultats suite à la prise en compte du modèle de turn-over

Nous pouvons alors calculer l'engagement au 31/12/2019 d'une part à l'aide des tables de turn-over « traditionnelles » explicitées dans la section précédente et d'autre part, à l'aide des modèles prédictifs du turn-over. La comparaison des résultats permet de se rendre compte de l'applicabilité de ces modèles dans le calcul de l'engagement et ainsi de conclure sur le modèle le plus pertinent. Les résultats sont répertoriés dans le tableau ci-dessous :

	VAP	PBO	SC
	214 901 886	96 165 190	6 555 668
GLM	237 295 769	101 486 537	7 079 304
CART	121 482 890	61 264 668	3 731 746
Random Forest	206 719 016	95 376 988	6 176 493
Gradient Boosting	212 566 752	97 023 442	6 364 975

TABLE 5.1 – Les résultats sur l'engagement des différents modèles de turn-over

On constate que l'engagement calculé à l'aide du Gradient Boosting présente le moins d'écart avec la VAP au 31/12/2019 calculée à l'aide des hypothèses classiques. Ce modèle semble d'une part offrir les meilleures performances mais il semble également s'approcher le plus des hypothèses classiques calibrées sur le même historique d'observations.

5.4 Les salaires projetés

Nous représentons les salaires projetés jusqu'au départ en retraite de nos salariés. Ces projections permettent de nous faire une idée des potentiels impacts de nos modèles sur l'engagement.

Les hypothèses classique d'évolution des salaires sont des taux de profil de carrière inflation comprise présentés en Annexe.7.

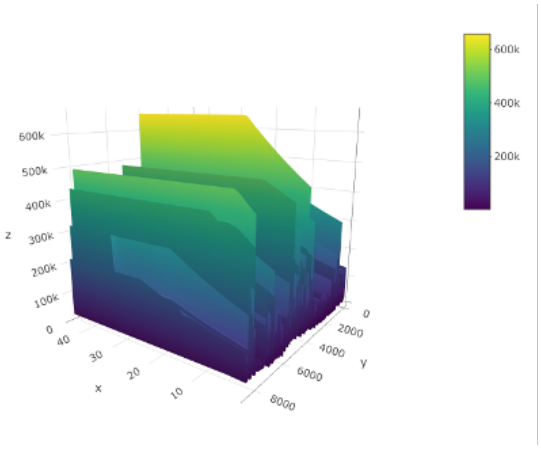


FIGURE 5.4 – Projections des salaires à l'aide des hypothèses classiques

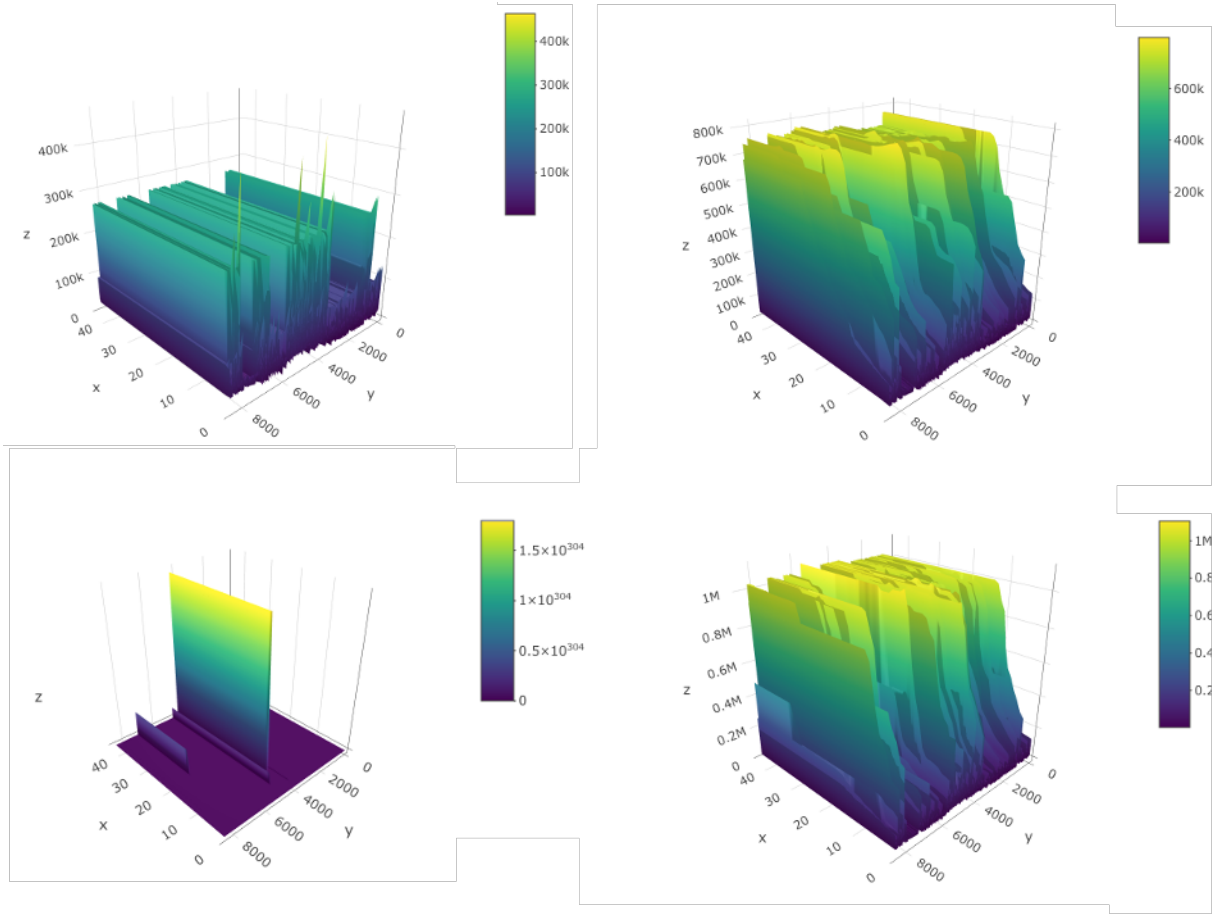


FIGURE 5.5 – Projections des salaires à l'aide des modèles prédictifs (modèles linéaires généralisés : en bas à gauche, arbre : en haut à gauche, random forest : en haut à droite et gradient boosting : en bas à droite)

Nous constatons que les salaires projetés par le modèle linéaire généralisé sont anormalement élevés. Le modèle dit régression log-gamma est très mal adapté à notre problématique, nous décidons donc de l'écarter dans ce qui suit.

5.5 Les résultats suite à la prise en compte du modèle d'évolution des salaires

Nous calculons comme pour le turn-over l'engagement au 31/12/2019 d'une part à l'aide des hypothèses « traditionnelles » explicitées dans la section précédente et d'autre part, à l'aide des modèles prédictifs. Les résultats sont répertoriés dans le tableau ci-dessous :

	VAP	PBO	SC
	214 901 886	96 165 190	6 555 668
CART	125 248 084	63 181 081	3 907 543
Random Forest	275 227 417	114 814 430	8 423 831
Gradient Boosting	395 411 391	134 221 102	11 725 110

TABLE 5.2 – Les résultats sur l'engagement des différents modèles d'évolution des salaires

On constate que l'écart en terme de PBO calculée à l'aide des hypothèses classiques est le plus faible à l'aide du modèle du Random Forest, nous le retenons donc finalement.

Nous remarquons toutefois que les écarts sont plus importants que ceux du turn-over, ceci s'explique par le fait que l'hypothèse dite « traditionnelle » d'évolution des salaires est d'une identique selon la catégorie socio-professionnelle, et de deux, constante durant la projection de carrière des salariés. Les différents modèles viennent donc prendre en compte différents paramètres augmentant ainsi l'engagement.

5.6 Les résultats suite à la prise en compte des promotions

Pour ce qui est des promotions, nous allons également comparer la PBO réelle avec la PBO estimées avec les différents algorithmes à la seule différence que l'impact sera moindre. En effet, l'hypothèse de promotion n'est pas prise en compte dans le calcul classique de la PBO puisque les taux d'augmentation des salaires dans le jeu d'hypothèses « traditionelles » ne différent pas selon la catégorie socio-professionnelle. L'impact se fera sur la prestation théorique lorsque le barème de droits différencie ceux-ci selon la catégorie socio-professionnelle mais également sur l'hypothèse de turn-over qui prévoit des tables différentes pour les cadres et non-cadres. Nous allons donc faire figurer les résultats de prise en compte des promotions selon les algorithmes ci-dessous.

	VAP	PBO	SC
	214 901 886	96 165 190	6 555 668
GLM	215 198 679	96 450 031	6 562 705
GLM sous-échantillonnage	214 156 673	96 364 059	6 534 908
GLM sur-échantillonnage	214 603 478	96 527 288	6 548 044
CART	214 174 775	96 104 499	6 536 754
CART sous-échantillonnage	214 415 784	96 006 973	6 540 837
CART sur-échantillonnage	213 755 818	96 202 523	6 525 926
Random Forest	214 924 924	96 175 602	6 556 009
Random Forest sous-échantillonnage	212 677 136	95 976 100	6 497 382
Random Forest sur-échantillonnage	214 868 680	96 163 850	6 554 396
Gradient Boosting	212 308 765	95 811 633	6 482 185
Gradient Boosting sous-échantillonnage	212 074 575	95 723 091	6 480 366
Gradient Boosting sur-échantillonnage	212 257 585	95 651 680	6 480 742

TABLE 5.3 – Les résultats sur l'engagement des modèles de prise en compte des promotions

On constate globalement un effet de réduction engendré par la prise en compte des promotions à l'aide de nos différents modèles qui s'explique, comme annoncé précédemment, par des taux de turn-over plus élevés chez les cadres que les non-cadres.

Pour ce qui est du choix final du modèle n'ayant pas base de comparaison étant donné que le modèle classique ne prend pas en compte les promotions, nous nous basons sur le critère AUC et sélectionnons la régression logistique.

5.6.1 Critiques des modèles

Les modèles ainsi construits permettent d'individualiser les hypothèses en fonction des caractéristiques propres à chacun des salariés en se basant sur les observations tirées de l'historique de 15 ans. Cet historique a initialement été choisi dans le but de constituer une base de données assez conséquente pour permettre l'apprentissage des modèles prédictifs et la prise en compte du plus grand nombre de variables explicatives.

Or, dans la pratique concernant les tables de turnover, celles-ci sont calibrées sur un historique moins important pouvant aller de 3 années jusqu'à 5 années de statistiques reflétant alors les tendances actuelles à la démission. On peut d'ailleurs faire référence au boxplot 3.1 où l'on constate que les taux de turnover diminuent sur les 15 années d'observation, c'est d'ailleurs une tendance que l'on constate au sein de l'intégralité des sociétés du périmètre étudié révèlant que globalement les salariés ont tendance à moins démissionner aujourd'hui qu'il y a 15 ans. Cette approximation doit étre nuancée car malgrè la sensibilité aux fluctuations conjoncturelles, notamment la crise économique de 2008 pouvant expliquer que les démissions se font plus rares on peut également penser que l'introduction de la rupture conventionnelle en 2008 dans le Code du travail vient se substituer aux démissions.

Pour ce qui des taux d'évolution des salaires le fait de calibrer cette hypothèse de façon rétrospective permet de prendre en compte certaines subtilités liées à l'âge du salarié, son

ancienneté mais ne permet pas d'intégrer la vision à moyen terme de la société quant à sa politique salariale ainsi que son budget accordé aux revalorisations des salaires par exemple.

Concernant la prise en compte des promotions, celle-ci aurait plus de sens lorsque l'hypothèse d'évolution future des salaires propose un taux différent selon la catégorie socio-professionnelle. En effet, l'impact du changement de barème de droit qui ne prévoit pas tout le temps un barème différencié pour les cadres que les non-cadres est minime. On constate d'ailleurs dans notre cas, que la réduction de l'engagement est en majeure partie expliquée par le turn-over qui prévoit des taux supérieurs pour les cadres. Néanmoins, à travers la prise en compte des promotions, notre volonté était de limiter les écarts actuarielles d'expérience nous allons donc procéder à une reconstitution de l'engagement entre le 31/12/2019 et le 31/12/2020 afin de voir si tel est bien le cas. Ci-dessous nous présentons l'évolution de l'engagement d'une part avec le modèle classique et d'autre part à l'aide du modèle de prise en compte des promotions.

	Modèle « classique »	Modèle promotions
PBO au 31/12/2019	**96 165 190**	**96 450 031**
Service Cost	6 555 668	6 562 705
Prestations théoriques	-3 871 575	-3 871 575
Interest Cost	786 124	788 400
PBO projetées	**99 635 407**	**99 929 562**
Ecarts actuariels d'expérience	**-5 128 299**	**-4 994 090**
PBO (hypothèses constantes)	**94 507 108**	**94 935 472**
Ecarts actuariels d'hypothèses	**+5 580 490**	**+5 607 677**
PBO au 31/12/2020	**100 087 598**	**100 543 149**

TABLE 5.4 – Evolution de l'engagement sociaux entre le 31/12/2019 et 31/12/2020

Le fait de prendre en compte les promotions dans le calcul des engagements permet donc de limiter les écarts actuariels d'expérience (5.1% de gain actuariel d'expérience sans la prise en compte des promotions et 4.9% de gain actuariel d'expérience avec), toutefois au vu des matrices de confusions présentent en Annexe.6, sur les promotions mises en évidence en 2020 sur l'effectif 2019 les modèles de prédictions n'en prédisent correctement que 57% en moyenne et commettent un nombre non négligeable d'erreurs de prediction.

5.7 Recalcul des passifs sociaux

Malgrè ces quelques points explicités dans la section qui précède, dans cette partie, nous appliquons les trois modèles simultanément afin de chiffrer l'impact global de leurs mises en place. Les calculs sont fait premièrement au 31/12/2019 puis au 31/12/2020 à l'aide d'une démographie à jour permettant d'appréhender l'applicabibilité de nos modèles dans le temps.

5.7.1 Engagements au 31/12/2019

Nous résumons l'impact de nos différents modèles sur l'engagement au 31/12/2019 dans le tableau ci-dessous :

	VAP	PBO	SC
PBO « classique » au 31/12/2019	**214 901 886**	**96 165 190**	**6 555 668**
Impact modèle promotions	+0.14%	+0,30%	+0,11%
Impact modèle salaires	+53.02%	+33.36%	+50.77%
Impact modèle turn-over	+0.83%	+3.91%	-2.29%
PBO finale au 31/12/2019	**330 920 781**	**132 292 200**	**9 740 516**

TABLE 5.5 – Engagements sociaux au 31/12/2019

5.7.2 Engagements au 31/12/2020

Nous calculons désormais l'engagement au 31/12/2020. Nous estimons tout d'abord la PBO projetée à l'aide de la formule 1.3.1, le calcul est présenté ci-dessous :

$$139\,209\,779 = 132\,292\,200 + 9\,740\,516 + 1\,092\,585 - 3\,915\,522$$

Nous calculons ensuite la PBO à hypothèses constantes au 31/12/2020 et la PBO de clôture au 31/12/2020 en utilisant les modèles prédictifs simultanément. Les résultats sont présentés ci-dessous :

	VAP	PBO	SC
PBO (hypothèses constantes)	**306 478 094**	**128 704 120**	**9 005 319**
PBO au 31/12/2020	**334 948 164**	**137 468 217**	**9 806 721**

TABLE 5.6 – Engagements sociaux au 31/12/2020

5.7.3 Evolution de l'engagement

Pour finir, nous présentons l'évolution de l'engagement à l'aide des élements calculés précédemment. La réconciliation est présentée ci-dessous :

	PBO
PBO au 31/12/2019	**132 292 200**
Service Cost	9 740 516
Prestations théoriques	-3 915 522
Interest Cost	1 092 585
PBO projetées	**139 209 779**
Ecarts actuariels d'expérience	**-10 505 659**
PBO (hypothèses constantes)	**128 704 120**
Ecarts actuariels d'hypothèses	**+8 764 097**
PBO au 31/12/2020	**137 468 217**

TABLE 5.7 – Evolution de l'engagement sociaux entre le 31/12/2019 et 31/12/2020 en utilisant les différents modèles

Conclusion

A travers cette étude, nous avons dans un premier temps ciblé les hypothèses ayant le plus d'impact sur l'engagement à l'aide d'une étude de sensibilité. Notre modélisation s'est donc portée sur le turn-over et les évolutions de salaires.

Aussi, nous avons, lors de la modèlisation, mis en balance deux approches des modèles paramétriques tels que les modèles linéaires généralisés et non-paramétriques à l'aide du *machine learning*. Cette opposition, bien que commune, s'est faite assez naturellement puisque adéquate à notre problématique. La modélisation des hypothèses de turn-over et d'évolution des salaires s'est avérée plus performante à l'aide de l'utilisation du *machine learning* et donc de l'approche non paramétrique, contrairement à la modèlisation des promotions où la régression logistique et l'approche paramétrique est nettement plus performante que le *machine learning*.

Chacune des deux approches, paramétrique ou non paramétrique, présente des avantages et inconvénients. L'approche non paramétrique présente l'avantage de ne pas émettre d'hypothèses quant à nos données en *inputs* bien que la phase de *hyperparameters tuning* peut être lourde en matière de temps de calcul. Alors que l'approche paramétrique est plus contraignante en termes d'hypothèses mais plus rapide à mettre en oeuvre.

Nous nous sommes également confronter au type de problème auquel nous faisons face. En effet, en fonction de la variable à expliquer mais surtout de l'*output* que l'on désire conserver, le problème peut prendre la forme d'une classification ou bien d'une régréssion.

En effet, nous avons pour le turn-over ainsi que la prise en compte des promotions fait face à une problématique de classification binaire et à toutes les problématiques qui en découlent notamment le déséquilibre. Bien que cette notion n'a été introduite que lors de l'implémentation des promotions, le déséquilibre est également présent dans la base de données du turn-over et sa correction n'a pas été nécessaire puisque nous conservions les probabilité d'affectation. Pour ce qui est de la correction du déséquilibre pour les promotions, l'approche algorithmique qui consite à ne pas dénaturer les données d'origine s'est révélée finalement plus performante que le ré-échantillonnage.

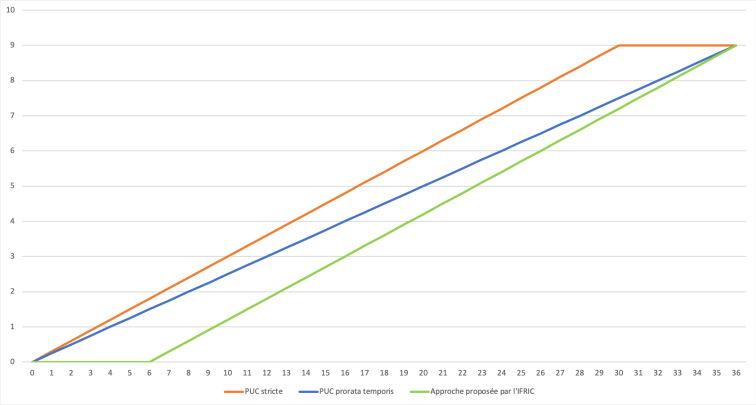
En somme, la création de modèles de prédiction permet de proposer des hypothèses adaptées aux caractéristiques individuelles des salariés et donc plus spécifiques que les hypothèses usuellement appliquées chez l'actuaire bien qu'à travers les chiffrages finaux ceci a un effet d'augmentation, dans notre cas, sur l'engagement.

Enfin, il convient de le rappeler, les résultats et les modélisations de cette étude sont adaptés à la méthode de calcul des passifs sociaux « PUC prorata temporis » qui est largement privilégiée et pour cause elle est recommandée par la norme IAS19. Cette étude mériterait d'être adaptée à la méthode nouvellement acceptée par l'IASB qui concerne les régimes suivants :

— les régimes où les salariés ont droit à une indemnité lorsqu'ils atteignent un certain âge de retraite à condition qu'ils soient employés par l'entité lorsqu'ils atteignent cet âge,

— les régimes où le montant de la prestation de retraite à laquelle un salarié a droit dépend de l'ancienneté du salarié avant l'âge de la retraite et est plafonné.

Autrement dit la majorité des régimes d'indemnité de fin de carrière est concernée. Cette nouvelle méthode consiste à linéariser les droits sur la période précédant l'obtention du « plafond » de droits et l'âge de départ en retraite. Nous pouvons résumer les différentes méthodes dans le graphique ci-dessous où la courbe verte représente les droits linéarisés à l'aide de la nouvelle méthode proposée.



Les différentes méthodes de calculs des engagements sociaux

A travers, l'exemple au Chapitre 1, nous avions montré la différence entre la méthode « PUC prorata temporis » et la méthode « PUC stricte », il conviendrait d'en faire de même avec cette fois-ci la nouvelle approche et ensuite d'adapter le modèle créé à l'occasion de cette étude, afin d'étudier les impacts des différents modèles prédictifs.

Bibliographie

[1] *International Accounting Standard 19 Employee Benefits*. IASB, 2011.

[2] Denis Clot. *Cours Analyses de données sous R*. ISFA, 2016.

[3] Esterina Masiello. *Cours Modèles linéaires généralisés*. ISFA, 2017.

[4] Ricco Rakotomalala. *Pratique de la Régression Logistique*. Université Lumière Lyon 2, 2017.

[5] Xavier Milhaud. *Cours Data Science en Actuariat (Big Data, Analytics)*. ISFA, 2018.

[6] Leo Breiman , Jerome Friedman, R.A. Olshen, and Charles Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[7] Leo Breiman. *Bagging predictors*. Machine learning, 1996.

[8] Leo Breiman. *Random Forests*. Machine learning, 2001.

[9] Robert E.Schapire and Yoav Freund. *A Short Introduction to Boosting*. Society for Artificial Intelligence, 1999.

[10] Jerome Friedman. *Greedy function approximation : A gradient boosting machine*. Annals of Statistics, 2001.

[11] Tian Qi Chen and Carlos Guestrin. *XGBoost : A Scalable Tree Boosting System*. University of Washington, 2016.

[12] INSEE, `https://www.insee.fr`

[13] Ricco Rakotomalala. *Analyse de corrélation*. Université Lumière Lyon 2, 2017.

[14] Wilkie. *A stochastic investment model for actuarial use* . , 1984.

[15] Christophe Bozetti and Billy Marques-Stenner. *Vers une baisse du passif social des indemnités de départ à la retraite*, Willi Towers Watson, 2021. `https://www.willistowerswatson.com`

[16] IFRS, `https://www.ifrs.org`

[17] Markit, `https://ihsmarkit.com`

Annexes

Annexe.1 Procédure de *tuning* du Gradient Boosting

Concernant le tuning du eXtreme Gradient Boosting ($xgboot$ sous R), nous procédons en utilisant la package $caret$ de R. En effet, la fonctionnalité $tuneGrid$ nous permet de spécifier un panel de paramètres à tester afin d'estimer les performances de notre modèle. Nous utilisons le critère AUC, pour rappel, fin de tester les modèles et conserver à chaque étape le plus performant.

La première étape consiste à tester un large panel de valeurs pour tous les paramètres du $xgboost$. Cette étape est la plus longue puisque correspond à environ 11 heures de temps de calcul pour la classification binaire aussi bien pour le turn-over que pour les promotions et 20 heures pour la régression pour la problématique des salaires. A noter toutefois que le fait de sur-échantillonner pour les modèles de promotions augmentent considérablement les temps de calcul.

A partir des résultats de la première étape, on peut affiner le *tuning* en faisant varier un hyper-paramètre à la fois et en fixant les autres selon les résultats de la première étape. Ainsi cette seconde étape est réalisée autant de fois qu'il n'y a d'hyper-paramètres et on sélectionne à chaque fois l'hyper-paramètre offrant la meilleure performance.

C'est qu'une fois cette procédure réalisée qu'on « entraine » le modèle finale par validation croisée sur l'échantillon composé de l'échantillon d'apprentissage et de test puis que l'on calcule les performances du modèle final sur un échantillon nouveau, l'échantillon de validation.

Annexe.2 Test de Shapiro & Wilk

Le test de Shapiro & Wilk permet de tester la normalité d'une série de données, il teste les hypothèses suivantes :

- H_0 : *« La distribution est normale »*
- H_1 : *« La distribution est non-normale »*

Ce test est basé sur la statistique suivante :

$$W = \frac{\left(\sum_i^n a_i \cdot x_{(i)}\right)^2}{\sum_i^n (x_i - \bar{x})^2}$$

où

— $x_{(i)}$ déisgne le i^{me} plus petit nombre dans l'échantillon,
— a_i sont des constantes générées à partir de la moyenne et de la matrice de variance covariance des quantiles d'un échantillon de taille n suivant la loi normale.

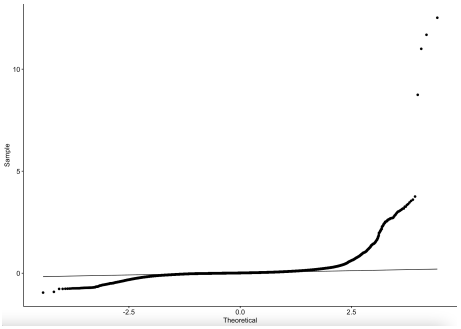
Sachant que l'hypothèse nulle est que la population est normalement distribuée :

— si la p-value est inférieure à un niveau α choisi (par exemple 0.05), alors l'hypothèse nulle est rejetée,
— au contraire, si la p-value est supérieure à α , alors on ne doit pas rejeter l'hypothèse nulle.

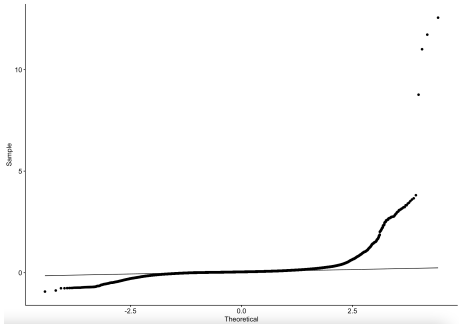
Dans notre cas, nous appliquons ce test aux $évolution\ des\ salaires$ d'une part et d'autre part l'$inflation$.

A savoir que ces deux différents tests n'impliquent pas la normalité du couple $(évolution\ des\ salaires, inflation)$. Il s'agit néanmoins d'une condition nécessaire. En effet, pour tester la normalité du couple $(évolution\ des\ salaires, inflation)$ il nous faut utiliser le test de multinormalité pour un vecteur aléatoire X de dimension n basé sur une modification du test de Shapiro-Wilk, présenté par Malkovitch et Afifi et reposant sur une approche vectorielle de l'étude de normalité. Ce test stipule que le vecteur X est normal si et seulement si toutes les combinaisons linéaires des composantes $(X_1, ..., X_n)$ de X sont normales. Dans notre cas, nous testons donc les combinaisons $évolution\ dessalaires - inflation$ et $évolution\ des\ salaires + inflation$.

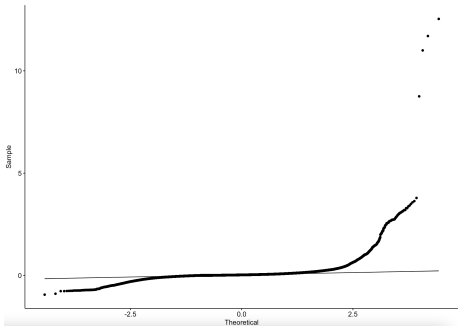
Toutefois, sous R le test de Shapiro est limité à un échantillon de 5000 données. Nous décidons donc de nous tourner plutot vers une analyse graphique en présentant les QQ-plots ci-dessous :



QQplot de la variable $évolution\ des\ salaires - inflation$



QQplot de la variable $évolution\ des\ salaires + inflation$



QQplot de la variable $évolution\ des\ salaires$

Annexe.3 Hyper paramètres des modèles d'évolution des salaires

Ci-dessous, nous présentons les hyper paramètres du random forest et gradient boosting suite au *tuning* des modèles de prédiction des salaires.

Annexe.3.1 *Hyperparameter tuning* random forest

$ntree$	$mtry$
500	8

Table 8 – Hyper-paramètres du random forest

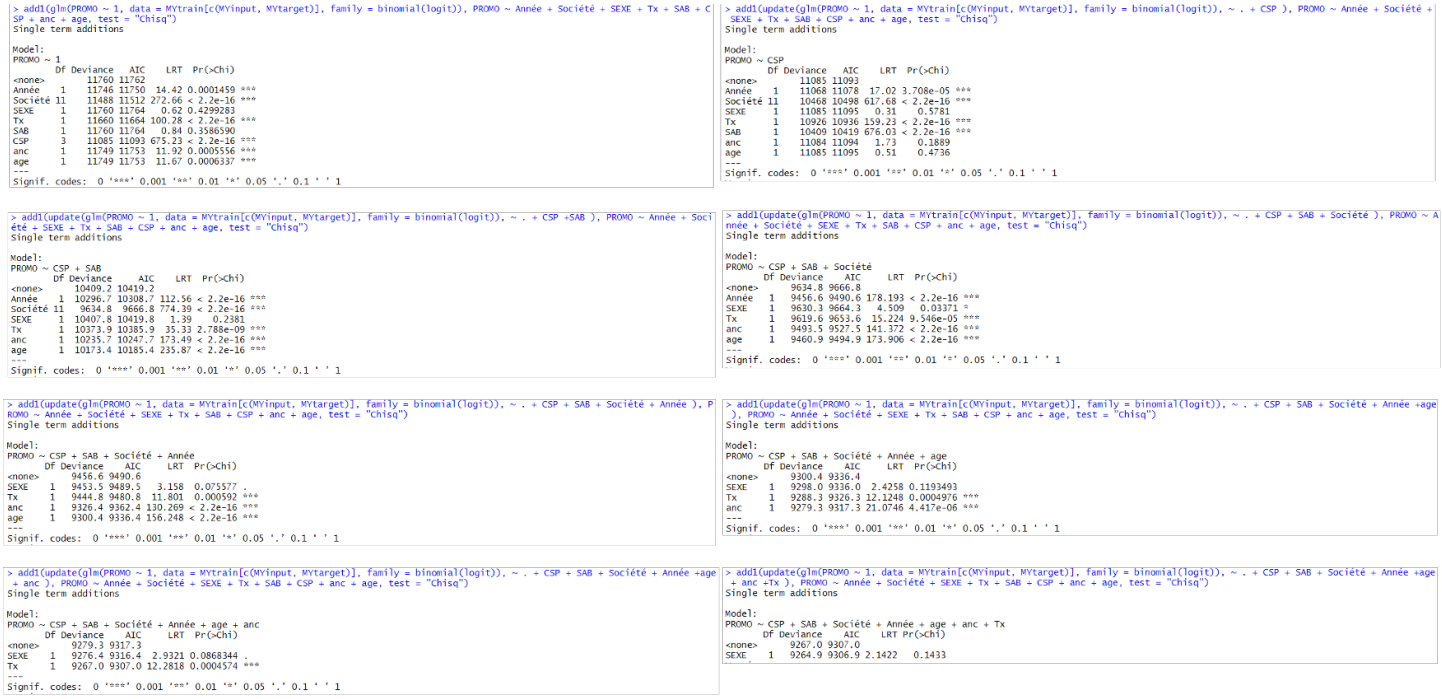
Annexe.3.2 *Hyperparameter tuning* gradient boosting

$nround$	max_depth	eta	$gamma$	$colsample\_bylevel$	min_child_weight	$subsample$
3000	6	0.02	0.7	0.6	3	1

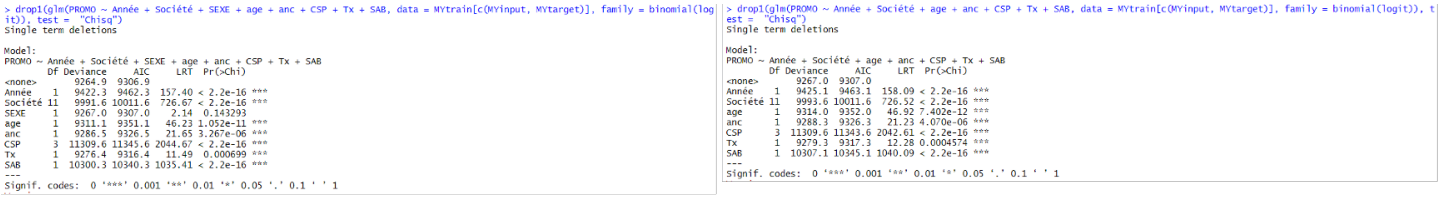
Hyper-paramètres du gradient boosting

Annexe.4 Sélection automatique des variables de la régression logistique appliquée à la modélisation des promotions

Nous présentons ci-dessous les résultats de la sélection automatique des variables lors de l'implémentation de la régression logitisque à notre problématique des promotions.



Résultats de la fonction $add1$ pour la modélisation des promotions



Résultats de la fonction $drop1$ pour la modélisation des promotions

Annexe.5 Hyper paramètres des modèles de prédcition des promotions

Ci-dessous nous présentons les hyper paramètres des différents random forest et gradient boosting suite au *tuning* des modèles de prédiction des promotions.

Annexe.5.1 *Hyperparameter tuning* random forest méthode de correction du déséquilibre avec optimisation du seuil

$ntree$	$mtry$
500	8

TABLE 9 – Hyper-paramètres du random forest pour les promotions

Annexe.5.2 *Hyperparameter tuning* random forest méthode de ré-échantilonnage over-sampling

ntree	*mtry*
500	8

Table 10 – Hyper-paramètres du random forest obtenu par *over-sampling*

Annexe.5.3 *Hyperparameter tuning* random forest méthode de ré-échantilonnage under-sampling

ntree	*mtry*
500	6

Table 11 – Hyper-paramètres du random forest obtenu par *under-sampling*

Annexe.5.4 *Hyperparameter tuning* gradient boosting méthode de correction du déséquilibre avec optimisation du seuil

nround	*max_depth*	*eta*	*gamma*	*colsample_bylevel*	*min_child_weight*	*subsample*
1000	3	0.05	0.05	1	1	1

Hyper-paramètres du gradient boosting pour les promotions

Annexe.5.5 *Hyperparameter tuning* gradient boosting méthode de ré-échantilonnage over-sampling

nround	*max_depth*	*eta*	*gamma*	*colsample_bylevel*	*min_child_weight*	*subsample*
800	2	0.1	0.1	1	1	0.75

Hyper-paramètres du gradient boosting obtenu par *over-sampling*

Annexe.5.6 *Hyperparameter tuning* gradient boosting méthode de ré-échantilonnage under-sampling

nround	*max_depth*	*eta*	*gamma*	*colsample_bylevel*	*min_child_weight*	*subsample*
2300	3	0.01	0.7	0.8	2	1

Hyper-paramètres du gradient boosting obtenu par *under-sampling*

Annexe.6 Matrices de confusion des modèles de prédiction des promotions

Optimisation du seuil		
Prédiction Réelle	Positive	Négative
Positive	49	37
Négative	692	7760

Sur-échantillonnage		
Prédiction Réelle	Positive	Négative
Positive	64	22
Négative	1201	7251

Sous-échantillonnage		
Prédiction Réelle	Positive	Négative
Positive	64	22
Négative	1153	7299

Prédiction Réelle	Positive	Négative
Positive	30	56
Négative	648	7804

Prédiction Réelle	Positive	Négative
Positive	70	16
Négative	1834	6618

Prédiction Réelle	Positive	Négative
Positive	64	22
Négative	2087	6365

Prédiction Réelle	Positive	Négative
Positive	1	85
Négative	11	8441

Prédiction Réelle	Positive	Négative
Positive	4	82
Négative	37	8415

Prédiction Réelle	Positive	Négative
Positive	67	19
Négative	1700	6752

Prédiction Réelle	Positive	Négative
Positive	53	33
Négative	851	7601

Prédiction Réelle	Positive	Négative
Positive	55	31
Négative	873	7579

Prédiction Réelle	Positive	Négative
Positive	63	23
Négative	1633	6819

TABLE 12 – Matrices de confusion associées à la régréssion logistique (première ligne), au CART (deuxième ligne), au Random Forest (troisième ligne) et au Gradient Boosting (dernière ligne)

Annexe.7 Hypothèses classiques de calcul des engagements sociaux de notre périmètre

Taux d'augmentation des salaires

Société	Taux d'augmentation salariale (inflation comprise)
2	3,2%
5	2%
9	2,5%
11	3%
14	2%
16	2,5%
17	1,8%
19	2,5%
21	3%
23	3%
24	3%
25	3%

Taux de profil de carrière en fonction des sociétés

Taux de turn-over

Société	2	5	9-Cadre	9-Non cadre	11	14	16-Etablissement 1	16-Etablissement 2	17-Cadre
16	2,50	12,03	25,63	22,11	6,51	9,37	0,20	45,95	14,51
17	2,50	12,03	25,63	22,11	6,29	9,37	0,19	44,01	14,51
18	2,50	12,03	25,63	22,11	6,07	9,37	0,18	42,11	14,51
19	2,50	12,03	25,63	22,11	5,80	9,37	0,18	40,25	14,51
20	2,50	12,03	25,63	22,11	5,58	9,37	0,17	38,44	14,51
21	2,38	11,50	25,63	22,11	5,36	8,85	0,16	36,66	13,75
22	2,26	10,99	23,83	22,11	5,14	8,33	0,15	34,93	13,04
23	2,15	10,50	22,17	22,11	4,92	7,92	0,14	33,24	12,40
24	2,03	10,03	20,61	22,11	4,76	7,50	0,14	31,59	11,75
25	1,87	9,59	19,17	22,11	4,54	6,98	0,13	29,98	11,16
26	1,80	9,15	17,83	22,11	4,32	6,56	0,12	28,41	10,58
27	1,68	8,73	16,58	20,56	4,16	6,14	0,12	26,89	9,99
28	1,56	8,34	15,42	19,13	3,94	5,73	0,11	25,41	9,40
29	1,48	7,95	14,34	17,79	3,77	5,42	0,10	23,96	8,81
30	1,37	7,58	13,34	16,54	3,61	5,00	0,10	22,56	8,28
31	1,29	7,22	12,40	15,38	3,39	4,69	0,09	21,21	7,81
32	1,21	6,87	11,54	14,31	3,23	4,27	0,09	19,89	7,34
33	1,09	6,54	10,73	13,31	3,06	3,96	0,08	18,62	6,87
34	1,01	6,21	9,98	12,37	2,90	3,65	0,08	17,38	6,35
35	0,94	5,89	9,28	11,51	2,74	3,33	0,07	16,19	5,88
36	0,86	5,58	8,63	10,70	2,57	3,02	0,07	15,04	5,46
37	0,82	5,27	8,03	9,95	2,46	2,81	0,06	13,93	5,11
38	0,74	4,98	7,46	9,26	2,30	2,50	0,06	12,87	4,70
39	0,70	4,70	6,94	8,61	2,19	2,29	0,05	11,84	4,29
40	0,62	4,42	6,46	8,01	2,02	1,98	0,05	10,86	3,94
41	0,59	4,15	6,00	7,45	1,91	1,77	0,04	9,92	3,53
42	0,55	3,89	5,58	6,92	1,75	1,56	0,04	9,02	3,23
43	0,47	3,63	5,19	6,44	1,64	1,35	0,04	8,16	2,88
44	0,43	3,38	4,83	5,99	1,53	1,15	0,03	7,35	2,64
45	0,39	3,13	4,49	5,57	1,42	1,04	0,03	6,57	2,35
46	0,35	2,89	4,18	5,18	1,31	0,83	0,03	5,84	2,12
47	0,27	2,65	3,88	4,82	1,20	0,73	0,02	5,15	1,82
48	0,23	2,42	3,61	4,48	1,09	0,62	0,02	4,50	1,59
49	0,16	2,19	3,36	4,17	0,98	0,52	0,02	3,89	1,41
50	0,12	1,97	3,12	3,87	0,93	0,42	0,01	3,33	1,18
51	0,08	1,75	2,91	3,60	0,82	0,31	0,01	2,80	1,06
52	0,08	1,54	2,70	3,35	0,77	0,21	0,01	2,32	0,88
53	0,04	1,33	2,51	3,12	0,66	0,10	0,01	1,88	0,71
54	0,04	1,12	2,34	2,90	0,60	0,10	0,01	1,48	0,59
55	0,00	0,92	2,17	2,70	0,55	0,00	0,00	1,12	0,47
56	0,00	0,73	2,02	2,51	0,44	0,00	0,00	0,81	0,35
57	0,00	0,53	1,88	2,33	0,38	0,00	0,00	0,53	0,24
58	0,00	0,34	1,75	2,17	0,33	0,00	0,00	0,30	0,12
59	0,00	0,15	1,63	2,02	0,27	0,00	0,00	0,11	0,06
60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Tables de turn-over en fonction des sociétés

Société	17-Non cadre	19	21-Cadre	21-Non cadre	23-Cadre	23-Non cadre	24-Cadre	24-Non Cadre	25
16	1,94	12,08	41,98	5,53	17,46	5,44	6,15	3,24	18,31
17	1,94	12,08	41,98	5,53	17,46	5,44	6,15	3,24	18,31
18	1,94	12,08	41,98	5,53	17,46	5,44	6,15	3,24	18,31
19	1,94	12,08	41,98	5,53	17,46	5,44	6,15	3,24	18,31
20	1,94	11,41	41,98	5,53	17,46	5,44	6,15	3,24	18,31
21	1,87	10,74	40,53	5,34	16,61	5,25	6,15	3,24	17,16
22	1,80	10,21	38,29	5,04	15,77	4,96	6,15	3,24	16,02
23	1,73	9,67	34,74	4,57	14,92	4,50	6,15	3,24	14,87
24	1,67	9,00	30,40	4,00	14,08	3,94	6,15	3,24	13,73
25	1,60	8,46	26,06	3,43	13,24	3,38	6,15	3,24	12,59
26	1,54	7,92	22,29	2,94	12,51	2,89	6,15	3,24	12,17
27	1,48	7,39	20,99	2,76	11,78	2,72	6,15	3,24	11,76
28	1,41	6,99	19,91	2,62	11,05	2,58	6,15	3,24	11,35
29	1,35	6,45	18,82	2,48	10,32	2,44	6,15	3,24	10,94
30	1,29	6,05	17,37	2,29	9,59	2,25	6,15	3,24	10,53
31	1,23	5,51	15,92	2,10	8,98	2,06	6,15	3,24	10,11
32	1,18	5,11	15,20	2,00	8,36	1,97	6,15	3,24	9,70
33	1,12	4,71	14,84	1,95	7,75	1,92	6,15	3,24	9,29
34	1,06	4,29	14,48	1,91	7,14	1,88	6,15	3,24	8,88
35	1,01	3,89	14,11	1,86	6,52	1,83	3,08	3,24	8,47
36	0,95	3,62	13,75	1,81	6,02	1,78	3,08	3,24	8,06
37	0,90	3,22	13,39	1,76	5,52	1,73	3,08	3,24	7,64
38	0,85	2,95	13,03	1,72	5,03	1,69	3,08	3,24	7,23
39	0,80	2,55	12,31	1,62	4,53	1,59	3,08	3,24	6,82
40	0,75	2,28	10,86	1,43	4,03	1,41	3,08	3,24	6,41
41	0,70	2,01	9,41	1,24	3,64	1,22	3,08	3,24	6,00
42	0,65	1,74	7,96	1,05	3,26	1,03	3,08	3,24	5,58
43	0,61	1,48	7,24	0,95	2,88	0,94	3,08	3,24	5,17
44	0,56	1,34	6,51	0,86	2,49	0,84	3,08	3,24	4,44
45	0,52	1,07	5,86	0,77	2,11	0,76	3,08	0,76	3,71
46	0,47	0,94	4,70	0,62	1,84	0,61	3,08	0,76	2,97
47	0,43	0,80	3,55	0,47	1,57	0,46	3,08	0,76	2,24
48	0,39	0,67	2,32	0,30	1,30	0,30	3,08	0,76	1,50
49	0,35	0,54	1,09	0,14	1,04	0,14	3,08	0,76	0,75
50	0,31	0,40	0,72	0,10	0,77	0,09	3,08	0,76	0,00
51	0,27	0,27	0,72	0,10	0,61	0,09	3,08	0,76	0,00
52	0,23	0,13	0,72	0,10	0,46	0,09	3,08	0,76	0,00
53	0,20	0,13	0,72	0,10	0,31	0,09	3,08	0,76	0,00
54	0,16	0,00	0,72	0,10	0,15	0,09	3,08	0,76	0,00
55	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
56	0,09	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
57	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
58	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
59	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Tables de turn-over en fonction des sociétés

Taux d'actualisation

Au 31/12/2019, le taux corporate bonds iBoxx AA 10+ était de 0,78%. Au 31/12/2020, le taux corporate bonds iBoxx AA 10+ était de 0,35%.

Taux de charges patronales

Société	Taux de charges patronales Cadre	Taux de charges patronales Non Cadre
2	46%	46%
5	56%	56%
9	43%	29%
11	45%	45%
14	44%	44%
16	50%	50%
17	57%	57%
19	47%	47%
21	51%	48%
23	51%	51%
24	50%	50%
25	46%	46%

Taux de profil de carrière en fonction des sociétés

Barèmes de droits

Ancienneté au terme	1	2-Cadre	2-Non cadre	3	5	6-Cadre	6-Non cadre
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1	0,00	0,00	0,13	0,00	0,00	0,00	0,00
2	0,50	0,00	0,25	0,00	0,00	1,00	0,10
3	0,50	0,00	0,38	0,00	0,00	1,00	0,15
4	0,50	0,00	0,50	0,00	0,00	1,00	0,20
5	1,00	0,75	0,63	1,50	0,00	1,00	0,25
6	1,00	0,90	0,75	1,50	0,00	1,00	0,30
7	1,00	1,05	0,88	1,50	0,00	1,00	0,35
8	1,00	1,20	1,00	1,50	0,00	1,00	0,40
9	1,00	1,40	1,13	1,50	0,00	1,00	0,45
10	2,00	1,60	1,25	2,50	1,00	2,00	1,00
11	2,00	1,80	1,42	2,50	1,10	2,00	1,10
12	2,00	2,00	1,58	2,50	1,20	2,00	1,20
13	2,00	2,20	1,75	2,50	1,30	2,00	1,30
14	2,00	2,45	1,92	2,50	1,40	2,00	1,40
15	2,00	2,70	2,08	3,00	1,50	2,00	1,50
16	2,00	2,95	2,25	3,00	1,60	2,00	1,60
17	2,00	3,20	2,42	3,00	1,70	2,00	1,70
18	2,00	3,45	2,58	3,00	1,80	2,00	1,80
19	2,00	3,70	2,75	3,00	1,90	2,00	1,90
20	3,00	3,95	2,92	4,00	2,00	3,00	2,00
21	3,00	4,20	3,08	4,00	2,10	3,00	2,10
22	3,00	4,45	3,25	4,00	2,20	3,00	2,20
23	3,00	4,70	3,42	4,00	2,30	3,00	2,30
24	3,00	4,95	3,58	4,00	2,40	3,00	2,40
25	3,00	5,20	3,75	4,50	2,50	3,00	2,50
26	3,00	5,45	3,92	4,50	2,60	3,00	2,60
27	3,00	5,70	4,08	4,50	2,70	3,00	2,70
28	3,00	5,95	4,25	4,50	2,80	3,00	2,80
29	3,00	6,20	4,42	4,50	2,90	3,00	2,90
30	4,00	6,45	4,58	5,00	3,00	4,00	3,00
31	4,00	6,70	4,58	5,00	3,10	4,00	3,10
32	4,00	6,95	4,58	5,00	3,20	4,00	3,20
33	4,00	7,20	4,58	5,00	3,30	4,00	3,30
34	4,00	7,45	4,58	5,00	3,40	4,00	3,40
35	5,00	7,50	4,58	6,00	3,50	4,00	3,50
36	5,00	7,50	4,58	6,00	3,60	4,00	3,60
37	5,00	7,50	4,58	6,00	3,70	4,00	3,70
38	5,00	7,50	4,58	6,00	3,80	4,00	3,80
39	5,00	7,50	4,58	6,00	3,90	4,00	3,90
40	6,00	7,50	4,58	7,50	4,00	4,00	4,00
41	6,00	7,50	4,58	7,50	4,10	4,00	4,10
42	6,00	7,50	4,58	7,50	4,20	4,00	4,20
43	6,00	7,50	4,58	7,50	4,30	4,00	4,30
44	6,00	7,50	4,58	7,50	4,40	4,00	4,40
45	6,00	7,50	4,58	7,50	4,50	4,00	4,50
46	6,00	7,50	4,58	7,50	4,60	4,00	4,60
47	6,00	7,50	4,58	7,50	4,70	4,00	4,70
48	6,00	7,50	4,58	7,50	4,80	4,00	4,80
49	6,00	7,50	4,58	7,50	4,90	4,00	4,90
50	6,00	7,50	4,58	7,50	5,00	4,00	5,00

Barèmes de droits en fonction des secteurs d'activité

Ancienneté au terme	7	9-Cadre	9-TAM	9-Emp/Ouv	10	11	12
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1	0,00	0,20	0,10	0,10	0,25	3,00	0,00
2	0,00	0,40	0,20	0,20	0,50	3,00	0,25
3	0,90	0,60	0,30	0,30	0,75	3,00	0,38
4	1,20	0,80	0,40	0,40	1,00	3,00	0,50
5	1,50	1,00	0,50	0,50	1,25	3,00	0,63
6	1,80	1,20	0,60	0,60	1,50	3,00	0,75
7	2,10	1,40	0,70	0,70	1,75	3,00	0,88
8	2,40	1,60	0,80	0,80	2,00	3,00	1,00
9	2,70	1,80	0,90	0,90	2,25	3,00	1,13
10	3,00	2,00	1,00	1,00	2,50	3,00	1,25
11	3,30	2,20	1,10	1,10	2,83	3,00	1,42
12	3,60	2,40	1,20	1,20	3,17	3,00	1,58
13	3,90	2,60	1,30	1,30	3,50	3,00	1,75
14	4,20	2,80	1,40	1,40	3,83	3,00	1,92
15	4,50	3,00	1,50	1,50	4,17	3,00	2,08
16	4,80	3,30	1,65	1,60	4,50	3,00	2,25
17	5,10	3,60	1,80	1,70	4,83	3,00	2,42
18	5,40	3,90	1,95	1,80	5,17	3,00	2,58
19	5,70	4,20	2,10	1,90	5,50	3,00	2,75
20	6,00	4,50	2,25	2,00	5,83	3,00	2,92
21	6,30	4,80	2,40	2,10	6,17	3,00	3,08
22	6,60	5,10	2,55	2,20	6,50	3,00	3,25
23	6,90	5,40	2,70	2,30	6,83	3,00	3,42
24	7,20	5,70	2,85	2,40	7,17	3,00	3,58
25	7,50	6,00	3,00	2,50	7,50	3,00	3,75
26	7,80	6,30	3,15	2,60	7,83	3,00	3,92
27	8,10	6,60	3,30	2,70	8,17	3,00	4,08
28	8,40	6,90	3,45	2,80	8,50	3,00	4,25
29	8,70	7,20	3,60	2,90	8,83	3,00	4,42
30	9,00	7,50	3,75	3,00	9,17	3,00	4,58
31	9,00	7,50	3,90	3,10	9,50	3,00	4,75
32	9,00	7,50	4,05	3,20	9,83	3,00	4,92
33	9,00	7,50	4,20	3,30	10,17	3,00	5,08
34	9,00	7,50	4,35	3,40	10,50	3,00	5,25
35	9,00	7,50	4,50	3,50	10,83	3,00	5,42
36	9,00	7,50	4,65	3,60	11,17	3,00	5,58
37	9,00	7,50	4,80	3,70	11,50	3,00	5,75
38	9,00	7,50	4,95	3,80	11,83	3,00	5,92
39	9,00	7,50	5,10	3,90	12,17	3,00	6,08
40	9,00	7,50	5,25	4,00	12,50	3,00	6,25
41	9,00	7,50	5,40	4,10	12,83	3,00	6,42
42	9,00	7,50	5,55	4,20	13,17	3,00	6,58
43	9,00	7,50	5,70	4,30	13,50	3,00	6,75
44	9,00	7,50	5,85	4,40	13,83	3,00	6,92
45	9,00	7,50	6,00	4,50	14,17	3,00	7,08
46	9,00	7,50	6,15	4,60	14,50	3,00	7,25
47	9,00	7,50	6,30	4,70	14,83	3,00	7,42
48	9,00	7,50	6,45	4,80	15,17	3,00	7,58
49	9,00	7,50	6,60	4,90	15,50	3,00	7,75
50	9,00	7,50	6,75	5,00	15,83	3,00	7,92

Barèmes de droits en fonction des secteurs d'activité

Tables de survie

A la date du 31/12/2019, les dernières tables de survie délivrées par l'INSEE était les tables INSEE 2013-2015.

Table des figures

1.1 Processus d'évaluation des passifs sociaux 15
1.2 Relation entre l'engagement et le service cost 17
1.3 Les différents éléments de la provision 22
1.4 Les différents éléments de la charge anuelle de retraite 22
1.5 Impact d'une réduction de régime sur l'engagement d'une société 25
1.6 Impact d'une liquidation de régime sur l'engagement d'une société 26
1.7 Récapitulatif de l'évolution de l'engagement d'une année à l'autre 27

2.1 Récapitulatif des hypothèses utilisées en pratique dans le calcul des indemnités de fin de carrière 29
2.2 Histogrammes de notre effectifs en fonction des années d'exercice 32
2.3 Pyramide des anciennetés (à gauche) et la pyramide des âges (à droite) 33
2.4 Histogramme des mouvements d'effectif en fonction des années 33
2.5 Histogramme des salaires moyens selon la catégorie socio-professionnelle 34
2.6 Impact de la variation du taux de charges patronales sur la VAPF (utilisation d'un barème par paliers) 37

3.1 Boxplot des taux de turn-over en fonction des années 42
3.2 Diagramme des valeurs propres 43
3.3 Graphique de corrélation des variables sur le plan engendré par les deux premiers axes 44
3.4 Double visualisation des variables et individus sur le plan engendré par les deux premiers axes 45
3.5 Procédé de la validation croisée K-fold 55
3.6 Résultat de la fonction *drop1* 56
3.7 Résultat de la fonction *add1* 57
3.8 Probabilités prédites à l'aide de la régression logistique entrainée sur l'*échantillon d'apprentissage* en fonction des variables explicatives 59

3.9 Moyenne des scores en fonction du nombre de démission sur l'*échantillon d'apprentissage* (à gauche) ainsi que sur l'*échantillon test* (à droite) 60
3.10 Arbre de décision maximal obtenu sur l'*échantillon d'apprentissage* 63
3.11 Arbre de décision élagué selon la règle « $1 - SE$ » 64
3.12 Erreur de prédiction sur l'*échantillon test* en fonction du nombre d'arbres . . . 65
3.13 Importance des variables dans la régression logistique (en haut à gauche), dans l'arbre CART (en haut à droite), le random forest (en bas à gauche) et le gradient boosting (en bas à droite) 69
3.14 Courbes ROC des différents modèles 70

4.1 Boxplot des taux d'évolution des salaires en fonction des années 73
4.2 Diagramme des valeurs propres 74
4.3 Graphique de contribution des variables selon les axes de l'ACP 74
4.4 Graphique de corrélation des variables sur le plan engendré par les deux premiers axes 75
4.5 Double visualisation des variables et individus sur le plan engendrépar les deux premiers axes 75
4.6 Inflation annuelle (en haut à droite), inflation mensuelle (en haut à gauche) et l'indice de prix à la consommation série 01763866 tirée de l'INSEE (en bas) . . 77
4.7 Densité des salaires annuels bruts 80
4.8 Résultat de la fonction *drop1* 81
4.9 Résultat de la fonction *add1* 81
4.10 Représentation graphique des résidus de Pearson du modèle 83
4.11 Importance des variables dans le GLM (en haut à gauche), dans l'arbre CART (en haut à droite), le random forest (en bas à gauche) et le gradient boosting (en bas à droite) 85
4.12 Boxplot du taux de promotion en fonction des années 87
4.13 Diagramme des valeurs propres 87
4.14 Graphique de corrélation des variables sur le plan engendré par les deux premiers axes 88
4.15 Double visualisation des variables et individus sur le plan engendré par les deux premiers axes 89
4.16 Histogramme de notre effectif en fonction des années 90
4.17 Courbe ROC et seuil optimal 91
4.18 Importances des variables de nos différents modèles de promotion (régréssion logistique : première ligne), CART : deuxième ligne, Random Forest : troisième ligne, Gradient Boosting : dernière ligne) 95
4.19 Courbes ROC des différents modèles 97

5.1 Fonctionnement de l'outil de calcul des IFC 99

5.2 Projections des probabilités du turn-over à l'aide des tables de turn-over classiques 99

5.3 Projections des probabilités du turn-over à l'aide des modèles prédictifs (régression logistique : en haut à gauche, arbre : en haut à droite, random forest : en bas à gauche et gradient boosting : en bas à droite) 100

5.4 Projections des salaires à l'aide des hypothèses classiques 101

5.5 Projections des salaires à l'aide des modèles prédictifs (modèles linéaires généralisés : en bas à gauche, arbre : en haut à gauche, random forest : en haut à droite et gradient boosting : en bas à droite) 102

Liste des tableaux

1.1 Tables de mortalité du moment (à gauche), générationnelles (à droite) 20

2.1 Constitution de l'échantillon pour la construction de la base de données 31
2.2 Impact de la variation du taux d'actualisation sur la VAPF 35
2.3 Impact de la variation du taux d'évolution des salaires (inflation comprise) sur la VAPF 36
2.4 Impact de la variation du taux de turn-over sur la VAPF 37

3.1 Base de données pour la modélisation du turn-over 40
3.2 Nombres d'axes à retenir pour l'ACP selon les critères de Kaiser et de Coude . 44
3.3 Récapitulatif du modèle du turn-over par régression logistique 55
3.4 Odds Ratio et Intervalles de confiance associés à la régression logistique entrainée sur l'*échantillon d'apprentissage* 58
3.5 Matrice de confusion 61
3.6 Critères de performance associés à la régression logistique obtenue par validation croisée sur l'*échantillon d'apprentissage et de test* 62
3.7 Critères de performance associés à l'algorithme CART obtenu par validation croisée sur l'*échantillon d'apprentissage et de test* 64
3.8 Hyper-paramètres du random forest obtenus par validation croisée sur l'*échantillon d'apprentissage et de test* 66
3.9 Critères de performance associés au random forest obtenu par validation croisée sur l'*échantillon d'apprentissage et de test* 67
3.10 Hyper-paramètres du gradient boosting obtenus par validation croisée sur l'*échantillon d'apprentissage et de test* 67
3.11 Critères de performance associés au gradient boosting obtenu par validation croisée sur l'*échantillon d'apprentissage et de test* 68
3.12 Les différents AUC calculés sur l'*échantillon de validation* selon les différents algorithmes 70

4.1 Base de données pour la modélisation de l'évolution des salaires 72

4.2 Nombres d'axes à retenir pour l'ACP selon les critères de Kaiser et de Coude . 74
4.3 Récapitulatif du modèle log-Gamma pour la modèlisation des salaires 82
4.4 Critères de performance associés à la régression log-Gamma 84
4.5 Critères de performance associés à l'algorithme CART 84
4.6 Critères de performance associés à l'algorithme Random Forest 84
4.7 Critères de performance associés à l'algorithme Gradient Boosting 85
4.8 Les différents critères selon les différents algorithmes 86
4.9 Base de données pour la modélisation des promotions 86
4.10 Nombres d'axes à retenir pour l'ACP selon les critères de Kaiser et de Coude . 88
4.11 Récapitulatif du modèle de régression logistique pour la modèlisation des promotions 92
4.12 Odds Ratio et Intervalles de confiance associés à la régression logistique entrainée sur l'*échantillon d'apprentissage et de test* 93
4.13 Critères de performance des régressions logistiques selon l'algorithme de correction du déséquilibre 93
4.14 Critères de performance des algorithmes CART selon l'algorithme de correction du déséquilibre 94
4.15 Critères de performance des algorithmes random forest selon l'algorithme de correction du déséquilibre 94
4.16 Critères de performance des algorithmes gradient boosting selon l'algorithme de correction du déséquilibre 94
4.17 Les différents AUC selon les différents algorithmes et la méthode de correction du déséquilibre utilisée 96

5.1 Les résultats sur l'engagement des différents modèles de turn-over 101
5.2 Les résultats sur l'engagement des différents modèles d'évolution des salaires . 103
5.3 Les résultats sur l'engagement des modèles de prise en compte des promotions 104
5.4 Evolution de l'engagement sociaux entre le 31/12/2019 et 31/12/2020 105
5.5 Engagements sociaux au 31/12/2019 106
5.6 Engagements sociaux au 31/12/2020 106
5.7 Evolution de l'engagement sociaux entre le 31/12/2019 et 31/12/2020 en utilisant les différents modèles 107

8 Hyper-paramètres du random forest 113
9 Hyper-paramètres du random forest pour les promotions 114
10 Hyper-paramètres du random forest obtenu par *over-sampling* 115
11 Hyper-paramètres du random forest obtenu par *under-sampling* 115
12 Matrices de confusion associées à la régression logistique (première ligne), au CART (deuxième ligne), au Random Forest (troisième ligne) et au Gradient Boosting (dernière ligne) 116