Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Imane ATMAN
Titre : Modélisation du taux d'incidence perte d'emploi avec les méthodes classiques et alternatives de Machine Learning

Confidentialité : ☐ Non ☑ Oui (Durée : ☐ 1 an ☑ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Olivier Lopez

Membres présents du Jury de la filière

Sébastien Farkas

Entreprise : Sogecap
Nom :
Signature : Mercè Batista

SOGECAP
Tour D2 - 17 bis place des Reflets
92919 Paris La Défense 2
Adresse Postale : TSA 61101
92919 Paris La Défense Cedex

Directeur de Mémoire en entreprise :
Nom : [tuteur entreprise] Ismail ELHADNI
Signature : i.Elhadni
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)

Secrétariat :

Bibliothèque :

Signature du responsable entreprise

Mercè Batista

Signature du candidat

Atman

Résumé

Ce mémoire se place dans le cadre de l'étude des *Cessione del Quinto*. Il s'agit d'un prêt particulier : les mensualités sont prélevées directement sur le salaire de l'emprunteur et elles ne peuvent excéder le cinquième celui-ci. Selon la loi italienne, ce prêt est obligatoirement couvert par une assurance pour les risque de décès de l'emprunteur ou bien d'insolvabilité due à une perte d'emploi, volontaire ou non.

En assurance, éviter les erreurs de financement constitue un objectif primordial pour assurer la pérénité des produits, la modélisation des risques est l'outil principal pour y parvenir. L'objet de ce mémoire porte sur l'étude actuarielle du produit CQS et l'identifiation des facteurs de risque auquel il est soumis.

Lors de notre démarche nous avons travaillé à partir des bases de données de la populations des assurés et des sinistrés dont nous disposons.
Dans un premier temps, nous avons calculé les taux d'incidence de l'évènement d'intérêt par année et selon la catégorie socio-professionelle. La méthode d'estimation des réserves de Chain Ladder, nous a permis d'estimer les IBNRs.

Les calculs selon la méthode de Kaplan Meier et le modèle de Cox nous ont apporté de nombreuses indications concernant le risque perte d'emploi. Ces méthodes sont habituellement utilisées en analyse de survie, car elle permettent la considération de la censure. L'application d'un lissage sur les taux bruts s'est montrée nécessaire.Nous avons utilisé la méthode de Whittaker-Henderson. Nous avons déduit les taux d'incidence perte d'emploi au cours du temps en fonction des différentes modalités.

L'étape finale de cette étude consiste à évaluer l'impact des différentes covariables, qui sont les caractéristiques des assurés, sur la variable cible, la survenance de la perte d'emploi, avec des méthodes de *Machine Learning*. Afin d'obtenir les meilleures prédictions possibles, plusieurs algorithmes ont été testés, via un processus très rigoureux et méthodique. Nous avons finalement retenu les résultats des méthodes les plus performantes dans le cadre de notre étude : les forêts aléatoires et le *Gradient Boosting*.

Nous distinguons les covariables les plus impactantes : la catégorie socio-professionelle, le sexe et le produit. Une approche de Machine Learning est un complément utile aux approches classiques pour effectuer une analyse plus poussée.
Cette étude est enrichissante et permet diverses applications actuarielles, telle que la pertinence et la cohérence du provisionnement.

Mots clés : Risque perte d'emploi, Analyse de Survie, Méthode de Kaplan Meier, Modèle de Cox, *Machine Learning*, Arbres de décision, Forêts aléatoires, *Gradient Boosting*

Abstract

This thesis is part of the *Cessione del Quinto* study. It is a particular loan : monthly payments shall be levied directly on the borrower's salary and shall not exceed one fifth of the borrower's salary. According to Italian law, this loan must be covered by insurance against the risk of death of the borrower or of insolvency due to loss of employment, whether voluntary or not.

In insurance, avoiding funding errors is a key objective for product continuity, and risk modelling is the main tool for achieving this. The purpose of this thesis is to do an actuarial analysis of the CQS product and to identify the risk factors to which it is subject.

We have been working from the databases of the populations of the insured and the claims we have. In the first step we calculated the interest event impact rates per year and by socio-professional category. The method of estimating Chain Ladder's reserves allowed us to estimate the IBNRs.

The Kaplan Meier method and Cox model impact rate calculations have given us many indications about the risk loss of employment. These methods are commonly used for survival analysis, as they allow censorship to be considered. The application of smoothing on gross rates was necessary.We used the Whittaker-Henderson method. We implied loss of employment incidence rates for different modalities.

The final step in this study is to assess the impact of different covariables, which are insured characteristics, on the target variable, the occurrence of loss of employment, with methods of Machine Learning. In order to obtain the best possible predictions, several algorithms were tested, through a very rigorous and methodical process. We finally use the results of the best methods for our study : random forests and the gradient boosting.

We distinguish the most important covariables : the socio-professional category, gender and product. A Machine Learning approach is a useful complement to traditional approaches, for further analysis. This analysis is enriching and enables various actuarial applications, such as the provisioning study.

Keywords : Loss of employment risk, Survival Analysis, Kaplan Meier Method, Cox Model, Machine Learning, Decision Trees, Random Forests, Gradient Boosting

Note de Synthèse

Cadre de l'étude

Le marché italien du Cessione del Quinto
CQS (*Cessione del Quinto dello Stipendio*) est un type particulier de prêt créé en 1914, permettant à l'emprunteur d'effectuer un prêt dont le montant maximal des mensualités ne peut être supérieur au cinquième de son salaire mensuel net ou de sa rente. L'une des particularité de ce prêt est que ses mensualités sont prélevées directement à la source par l'employeur et versées tous les mois au créancier. La durée totale du prêt ne peut excéder 10 ans et n'est généralement pas inférieure à 24 mois.

De nouvelles normes ont autorisé à étendre l'octroi de ces prêts CQS aux employés du secteur privé ainsi qu'aux titulaires d'une rente (retraités et invalides), distinguant ainsi deux types de produits :

— CQS (*Cessione del Quinto dello Stipendio*) pour les actifs,

— CQP (*Cessione del Quinto della Pensione*) pour les retraités et titulaires d'une rente d'invalidité

Nous nous concentrons désormais sur les CQS, destinées aux actifs, le produit CQP étant réservé aux retraités.

Dans le cas du CQS uniquement, il existe un prêt complémentaire, appelé *Delega*, qui consiste en un second prêt additionnel au CQS ne présentant pas les mêmes caractéristiques.

Pour les employés du secteur privé et para-publique uniquement, le TFR (*Trattamento di Fine Rapporto*) est l'indemnité de départ, régie par le code du travail italien . Lorsqu'une rupture du contrat de travail est à l'initiative de l'employeur ou non, ce dernier a l'obligation de verser une indemnité de départ au salarié. Celle-ci comprend entre autres les congés annuels accumulés non utilisés, le paiement de la période de préavis obligatoire, une aide à la recherche d'un nouvel emploi... Le TFR correspond à environ 1 mois de revenu brut par année d'ancienneté. Il alimente un compte "épargne" dont l'assuré pourra bénéficier lorsqu'il quittera l'entreprise.

Les contrats Cessione del Quinto commercialisés par SOGECAP/SOGESSUR
En cas de rachat du prêt par l'emprunteur : Dans le cadre d'un remboursement anticipé ou de transfert du prêt, la couverture d'assurance cessera d'être effective et l'assureur remboursera une partie de la prime liée proportionnellement à la période d'assurance non prise en charge.

En cas de décès de l'emprunteur : Dans le cadre d'un décès, il est obligatoire de fournir un certificat

de décès et des documents pour l'organisme de crédit. Alors, l'assureur verse à la société octroyant le prêt le montant du capital restant dû à la date du décès, calculé à partir du montant du prêt, y compris les intérêts et les frais. Le taux annuel nominal (TAN) est mis en place par l'organisme de crédit au moment du prêt, pour permettre de calculer les intérêts.

En cas de perte définitive d'emploi : Dans le cadre d'un licenciement ou d'une démission, avant de solliciter l'assureur, l'emprunteur doit respecter une série d'obligations visant à recouvrer sa dette. Il doit aussi transmettre les documents nécessaires pour vérifier l'existence et le montant actuel du remboursement résiduel.

Données disponibles et statistiques descriptives

Description des bases de données (population des assurés et population des sinistrés)
Nous d'une base de données, regroupant celle des assurés et des sinistrés. Les valeurs aberrantes et incohérences numériques ont été traitées. Les variables explicatives que nous utilisons sont :

- le sexe de l'emprunteur,
- la catégorie socio-professionnelle,
- l'âge à la souscription,
- la date de souscription,
- le montant emprunté,
- la durée de l'emprunt,
- le type d'emprunt.

La variable que nous cherchons à modéliser est la sinistralité.

Analyse univariée
Nous commençons avec des statistiques descriptives de nos variables, qui indiquent les effectifs des variables qualitatives et, pour les variables numériques, des indicateurs de centralité et de dispersion comme la moyenne et les quartiles.

Analyse bivariée : corrélation entre les variables
Dans cette étape, nous cherchons à croiser nos variables explicatives et à mesurer leur corrélation. En effet, notre choix de modélisation implique l'absence de liaisons trop fortes entre les variables explicatives.

Deux coefficients sont utilisés afin de déterminer la relation entre deux variables quantitatives : le coefficient de Pearson et le coefficient de Spearman.

D'après les tableaux suivants, il n'y a aucune corrélation très forte entre les variables explicatives quantitatives.

	Age	Insured_amount	Contract_period	TFR_multiplicator	TFR_at_subscription
Age	1.0	0.25	0.021	-0.35	-0.33
Insured_amount	0.25	1.0	0.62	-0.38	-0.27
Contract_period	0.021	0.62	1.0	-0.12	-0.058
TFR_multiplicator	-0.35	-0.38	-0.12	1.0	0.77
TFR_at_subscription	-0.33	-0.27	-0.058	0.77	1.0

FIGURE 1 – Matrice de corrélation de Spearman

	Age	Insured_amount	Contract_period	TFR_multiplicator	TFR_at_subscription
Age	1.0	0.27	0.063	-0.33	-0.11
Insured_amount	0.27	1.0	0.56	-0.34	0.018
Contract_period	0.063	0.56	1.0	-0.11	0.072
TFR_multiplicator	-0.33	-0.34	-0.11	1.0	0.47
TFR_at_subscription	-0.11	0.018	0.072	0.47	1.0

FIGURE 2 – Matrice de corrélation de Pearson

Pour étudier les liaisons entre nos variables qualitatives, nous utilisons le V de Cramer. Comme l'indique la matrice ci-après, les liaisons entre les variables explicatives sont négligeables.

	Sex	Name_of_partner	Type_of_product	Socio_professional_category
Sex	1.0	0.0	0.0	0.03
Name_of_partner	0.0	1.0	0.01	0.01
Type_of_product	0.0	0.01	1.0	0.02
Socio_professional_category	0.03	0.01	0.02	1.0

FIGURE 3 – Matrice de Cramer

Première étude du taux d'incidence chômage

Désormais nous calculons le taux d'entrée annuel au chômage de notre portefeuille selon deux granulations : le partenaire financier et la catégorie socio-professionnelle des assurés en portefeuille.

	PUB	PARA	PRIV	Total
2015	0,00%	0,00%	0,00%	0,00%
2016	0,11%	1,60%	2,38%	0,83%
2017	0,25%	0,69%	3,37%	1,43%
2018	0,23%	0,49%	2,89%	1,33%
2019	0,00%	0,00%	0,00%	0,00%
Total	0,19%	0,40%	2,00%	0,93%

TABLE 1 – Taux d'incidence chômage sur l'ensemble du portefeuille

Le CQS est un produit en plein essor et très apprécié pour sa forte rentabilité due à des taux de sinistralités faibles.
Nous remarquons que le taux est environ 10 fois plus élevé pour le secteur privé que pour les fonctionnaires. C'est un constat tout à fait cohérent.

Avec la méthode de Clain ladder, nous allons estimer les IBNR sur notre période d'observation, à savoir survenus avant le 01 décembre 2019 mais non déclarés au 01 décembre 2019. Nous recalculons alors les taux d'entrée annuels au chômage.

	PUB	PARA	PRIV	Total
2015	0,00%	0,00%	0,00%	0,00%
2016	0,11%	1,60%	2,38%	0,83%
2017	0,25%	0,69%	3,37%	1,43%
2018	0,68%	0,98%	3,02%	1,65%
2019	0,49%	0,90%	2,65%	0,45%
Total	0,47%	0,94%	2,86%	1,46%

TABLE 2 – Taux d'incidence chômage sur l'ensemble du portefeuille (en considérant les IBNRs)

Les taux d'incidence sur 2019 semblent bien plus alignés à ceux des années précédentes.

Modélisation avec les méthodes d'analyse de survie

Les modèles non paramétriques L'estimateur de Kaplan Meier est l'un des principaux estimateurs non paramétriques de la fonction de survie. L'avantage important de cet estimateur est qu'il permet de prendre en compte les données censurées. L'estimateur de la fonction de survie Kaplan Meier est :

$\hat{S}(t) = \prod_{i=1,T_i<t}^{n} [\frac{n-i}{n-i+1}]^{\delta_i}$

A partir de l'estimateur de la fonction de survie, nous pouvons aisément déduire les taux bruts de mortalité via la relation :

$\hat{q}_x = 1 - \prod_{i=a_1}^{a_m} \frac{r_i - d_i}{r_i}$

Les modèles semi-paramétriques Le modèle de Cox est la méthode d'analyse de survie la plus couramment utilisée parmi les modèles semi-paramétriques. Il permet d'exprimer la fonction de risque instantané λ en fonction du temps t et des variables explicatives $X_1, \ldots, X_n$.

Nous avons alors $\lambda(t, X_1, \cdots, X_n) = \lambda_0(t) \exp(\Sigma_{i=1}^n \beta_i X_i)$ avec : $\beta = (\beta_1, \ldots, \beta_n)$ le vecteur des paramètres et $X = (X_1, \ldots, X_n)$ le vecteur des n variables explicatives (pouvant d'ailleurs dépendre du temps).

Dans cette formule, $\lambda_0(t)$ est appelé la fonction de hasard de base. Il correspond au risque instantané de décès lorsque toutes les variables sont nulles.
La deuxième partie $\exp(\Sigma_{i=1}^n \beta_i X_i)$ ne dépend quant à elle que des variables.
Ce modèle est appelé semi-paramétrique, en effet, nous ne cherchons pas à estimer $\lambda_0(t)$, mais le rapport des risques instantanés de décès pour des individus exposés différemment.

Il découle de cette formule une hypothèse primordiale pour le modèle de Cox : l'hypothèse de proportionnalité des risques et l'hypothèse de log linéarité.

Nous estimons les paramètres du modèle avec la méthode du maximum de vraissemblance. Il est nécessaire de vérifier les hypothèses, afin de s'assurer de la fiabilité des résultats.

Les méthodes de lissage Après avoir estimé les taux bruts de survenance du chômage dans notre portefeuille, qui présentent certaines inégalités dues à l'imperfection des conditions de l'expérience ; il est nécessaire de procéder à un lissage de nos valeurs brute afin de présenter de manière plus fidèle la loi que nous souhaitons estimer.

Nous distinguons deux types de modèles de lissages. Les modèles paramétriques telle que la méthode de Spline sont basés sur l'utilisation d'une loi sous-jacente usuelle à déterminer. Les méthodes non paramétriques, telles que les moyennes mobiles ou la méthode de Xhittaker-Henderson, quant à elles, ne reposent sur aucune famille de loi. Afin de déterminer le modèle de lissage le plus adapté il existe de nombreuses métriques telles que le coefficient de détermination R^2, l'erreur absolue moyenne en pourcentage (MAPE), les tests d'adéquation et de Kolmogorov Smirnv.

Applications Nous avons comparé, pour chacun des graphiques et pour chacune des méthodes de lissages, divers critères de validations. Nous sélectionnons le lissage le plus adapté : la méhode de Whittaker Henderson.

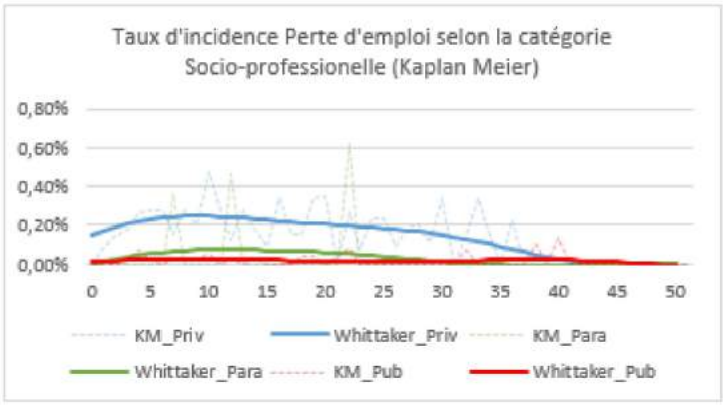


FIGURE 4 – Taux d'incidence perte d'emploi selon la catégorie Socio-professionelle (KM)

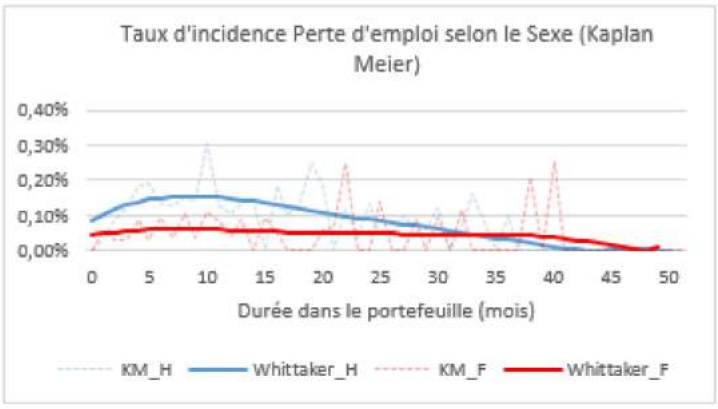


FIGURE 5 – Taux d'incidence Perte d'emploi selon le Sexe (KM)

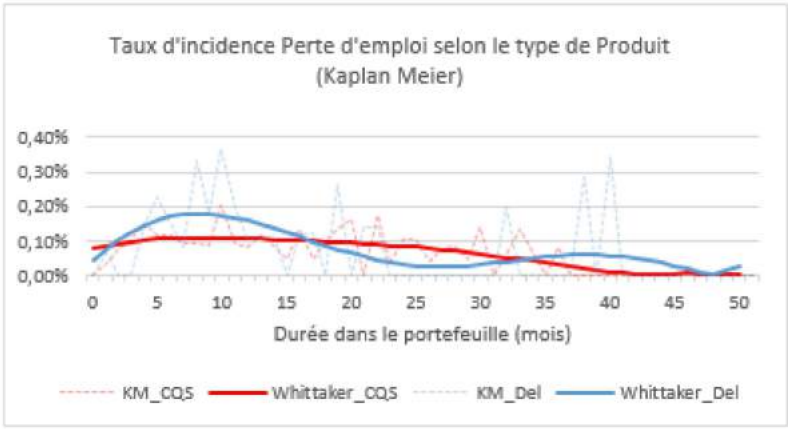


FIGURE 6 – Taux d'incidence Perte d'emploi selon le type de produit (KM)

Les statistiques sommaires ci-dessous indiquent l'importance des covariables dont nous disposons dans la prévision du risque perte d'emploi obtenu avec le modèle de Cox. Ainsi, d'après ce modèle,

les deux principales variables sont la catégorie socio-professionnelle et le nom du partenaire. Le type de produit, le sexe et le montant assuré se positionnent ensuite. Enfin, l'âge et la durée du contrat semblent de faible importance.

model	lifelines.CoxPHFitter
duration col	'Duration'
event col	'Claim'
baseline estimation	breslow
number of observations	10087
number of events observed	177
partial log-likelihood	-1429.31
time fit was run	2020-05-07 11:28:10 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
Age	-0.02	0.98	0.01	-0.04	-0.00	0.96	1.00	-2.15	0.03	4.89
Insured_amount	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-2.15	0.03	4.89
Contract_period	-0.01	0.99	0.00	-0.01	0.00	0.99	1.00	-1.79	0.07	3.77
Sex_M	0.52	1.68	0.16	0.19	0.86	1.17	2.40	2.82	<0.005	7.66
Name_of_partner_OTHERS	-0.31	0.73	0.37	-1.04	0.42	0.35	1.52	-0.83	0.40	1.31
Type_of_product_DELEGA	0.60	1.83	0.20	0.22	0.99	1.24	2.69	3.06	<0.005	8.83
Socio_professional_category_PRIV	1.37	3.95	0.59	0.22	2.53	1.25	12.53	2.34	0.02	5.66
Socio_professional_category_PUB	-0.59	0.55	0.62	-1.80	0.63	0.16	1.88	-0.94	0.35	1.53

Concordance	0.80
Log-likelihood ratio test	218.66 on 8 df
-log2(p) of ll-ratio test	139.96

FIGURE 7 – Paramètres du modèle de Cox

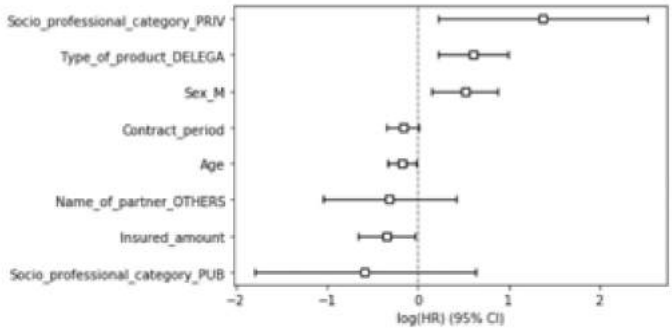


FIGURE 8 – Importance des covariables du modèle de Cox

Finalement, le modèle de Cox nous permet d'obtenir les fonctions de survie de chacun des individus.

Nous obtenons, en accord avec nos résultats précédents, en analysant les profils des individus sélectionnés, les conclusions suivantes : Les individus de la population à risque sont :

— des employés du secteur privé,
— des emprunteurs du produit *Delega*,
— des hommes,
— financés par le Partenaire1,
— d'âges très variables,
— emprunteurs de montants très différents sur des durées variables.

Les individus de la population moins risquée sont :

— employés du secteur public,
— des emprunteurs d'un produit CQS,

— des femmes,
— financés par un autre partenaire,
— d'âges très variables,
— emprunteurs de montants très différents sur des durées variables

Machine Learning

Généralités sur le Machine Learning Le principe du *machine learning*, est d'apprendre à l'ordinateur à partir des données. Nous effectuons un découpage de la base en deux parties, c'est l'échantillonnage. Les modèles d'apprentissage seront appliqués sur l'échantillon d'apprentissage, appelé *train* et le modèle qui ressortira sera testé sur l'échantillon d'entraînement appelé *test*.

Les arbres de décision : CART Ce sont des modèles simples qui permettent d'obtenir des résultats intuitifs et facilement compréhensibles, représentés sous forme graphique d'un arbre.
Un arbre de décision est une structure hiérarchisée formée de nœuds. Chaque nœud se divise au maximum en deux nœuds, un droit et un gauche.

A chacun des nœuds, une variable d'entrée est choisie et répartie en deux groupes.

Nous utilisons les arbres de survie. Ils sont une forme d'arbres de décisions adaptée pour traiter les données d'analyse de survie. L'arbre de décision se construit en une partition récursive des données en fixant un seuil pour chaque variable, mais elle ne peut pas rendre en compte les interactions entre les variables ni les informations censurées dans le modèle. Les critères utilisés pour les arbres de survie peuvent être regroupés en deux catégories :
— Maximiser l'hétérogénéité entre les nœuds,
— Minimiser l'homogénéité à l'intérieur des nœuds.
Néanmoins, les arbres de décisions sont sensibles au risque de sur-apprentissage.Les méthodes d'aggrégation sont des alternatives permettant de limiter ce risque.

Le Bagging - Les forêts aléatoires C'est un algorithme de classification, permettant de réduire la variance des prévisions d'un seul arbre de décision. Cet algorithme effectue un apprentissage sur plusieurs arbres de décisions construits aléatoirement et entraînés sur des sous-ensemble de données légèrement différents les uns des autres.

Pour cela, il effectue une approche de type *bagging*. Il permet de réduire la variance associée à l'estimateur et donc améliorer la stabilité et la performance des arbres de décision. Pour les Forêts aléatoires nous disposons donc d'un échantillon d'apprentissage $Z = \{(x_1, y_1), ..., (x_n, y_n)\}$. $X_1, \ldots, X_p$ sont les p variables explicatives. En construisant B arbres aléatoires dans la forêt, le Bagging effectue B tirages aléatoires avec remise parmi Z. L'algorithme crée ensuite un arbre sur chaque échantillon, en réalisant, pour la construction de chaque noeud un tirage de m variables parmi les p variables disponibles afin former la décision associée au noeud.
La prédiction correspond alors à la classe majoritaire.

Le nombre optimum d'arbre, pouvant être très nombreux, est un paramètre important, fonction des données et variables du problème. Nous l'obtenons en analysant l'erreur OOB en fonction du nombre

d'arbres utilisés.

Les *Random Survival Forest* sont une extension des forêts aléatoires classiques, adaptée aux données censurées.

La méthode générale est la même, néanmoins, au moment de la division des nœuds des arbres en deux sous-ensembles, nous utilisons un critère approprié pour les données censurées, le test du log rank.

Le Boosting - XG Boost Les algorithmes de Boosting sont des algorithmes itératifs de descente de gradients fonctionnels. Ils ont pour but de trouver les valeurs optimales des paramètres d'une fonction donnée. L'idée se rapproche de celle du bagging qui, plutôt que d'utiliser un seul modèle, en utilise plusieurs, qui sont ensuite agrégés, afin d'obtenir un unique et meilleur résultat.

L'intérêt de cette méthode est de réduire la variance, tout comme le bagging, mais aussi le biais de la prévision.

Evaluer les performances des modèles de machine learning Un bon modèle de *machine learning* doit permettre de généraliser, à savoir être capable de faire des prédictions à partir des données utilisées pour mettre en place le modèle mais surtout à partir de nouvelles données. La matrice de confusion permet de déterminer la qualité d'un algorithme de classification.

Applications En répartissant nos données en deux classes, la population de sinistrés et la population de non sinistrés nous obtenons alors des classes déséquilibrées. Pour y remédier , nous allons utiliser l'échantillonnage ascendant et réechantilloner nos données.

Les arbres de survie Les arbres de survie permettent de répartir la population en classes, selon les caractéristiques des individus.

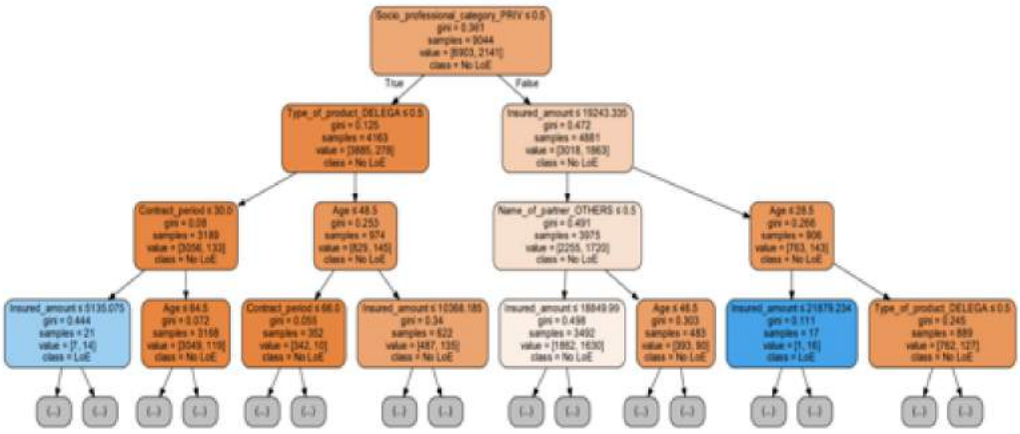


FIGURE 9 – Arbre de survie de notre portefeuille

Deux catégories sont identifiées comme population à risque.

Nous trouvons ainsi :

— Les individus du secteur privé dont le montant assuré est supérieur à 19 000 euros et l'âge inférieur à 28 ans.

— Les individus du secteur public ayant effectué un prêt *Delega* dont la durée est inférieure à 30 mois.

Les autres emprunteurs ne semblent pas présenter des risques importants et significatifs. Nous rappelons que cet arbre reste très instable et dispose de faibles capacités de prédiction. Ainsi, nous préférerons les méthodes d'agrégation qui suivent.

Les forêts aléatoires de survie A partir de la matrice de confusion du modèle, nous calculons la précision, le rappel et le f1-score qui fournissent d'excellents résultats.

```
[[2975   32]
 [   0  869]]
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      3007
           1       0.96      1.00      0.98       869

    accuracy                           0.99      3876
   macro avg       0.98      0.99      0.99      3876
weighted avg       0.99      0.99      0.99      3876
```

FIGURE 10 – Résultats obtenus avec les forêts aléatoires

En effet, les valeurs des indicateurs sont entre 0,96 et 1. Ainsi, le modèle, lors de la vérification sur la base d'entraînement, a obtenu des résultats très proches de la réalité. La moyenne des indicateurs est de 0,99.

Le Gradient Boosting Au vu de la qualité du modèle d'agrégation précédent, nous choisissons d'en implémenter un second type, le *Gradient Tree Boosting*. C'est un algorithme beaucoup plus complexe que le précédent et très coûteux en temps de calcul. Il a été très long a mettre en place dans notre outil.

A partir de la matrice de confusion du modèle, nous calculons les indicateurs.

```
[[2961   46]
 [   0  869]]
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      3007
           1       0.95      1.00      0.97       869

    accuracy                           0.99      3876
   macro avg       0.97      0.99      0.98      3876
weighted avg       0.99      0.99      0.99      3876
```

FIGURE 11 – Résultats obtenus avec le Gradient Boosting

Nous obtenons, une fois encore, d'excellents indicateurs.

L'importance des variables Les modèles agrégés ont une meilleure précision et capacité de prédiction. Ainsi ils expliquent le lien entre les variables explicatives et la variable cible. Bien que ces modèles soient difficilement interprétables, il est possible d'extraire des modèles agrégés l'influence de chaque variable.

Sur l'ensemble du portefeuille, la variable la plus importante est clairement la catégorie socioprofessionnelle. Les employés du secteur privé sont une population plus à risque que les employés du secteur public ou para-public.
Le nom du partenaire est la seconde variable la plus influente. Les prêts financés par le Partenaire1 semblent bien plus risqués que les autres partenaires. Néanmoins, étant donné le très faible historique dont nous disposons sur les autres partenaires, nous ne considérons pas ce résultat comme pertinent.
Le montant assuré intervient ensuite négativement, plus il est élevé moins le risque ne l'est.
Le sexe a aussi une influence non négligeable, la population masculine est plus à risque que la population féminine. Nous rappelons qu'une distinction tarifaire sur le sexe n'est pas autorisée, mais il est tout de même très intéressant de le remarquer, afin de rester vigilant quant à la proportion d'hommes dans le portefeuille.
La variable type de prêt nous indique que les prêts CQS ont une influence positive contrairement aux prêts secondaires, les *Delega*, qui augmentent le risque.
La durée du contrat et l'âge arrivent en dernière position. Ils influent négativement, plus ils sont élevés, moins l'assuré est à risque de Perte d'emploi.

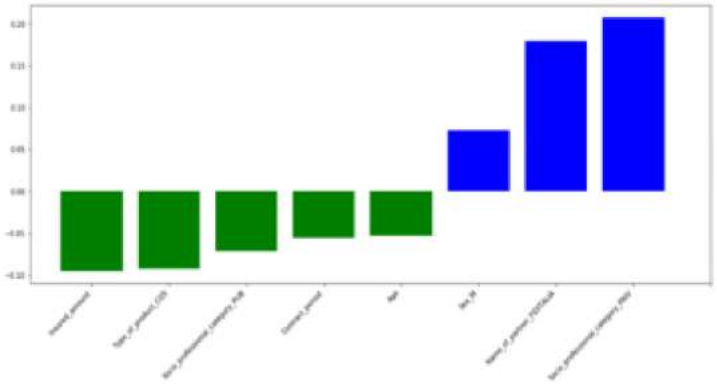


FIGURE 12 – Importance des variables sur l'ensemble du portfeuille

Afin d'affiner nos résultats nous effectuons une étude semblable en divisant le portefeuille en trois catégories selon le secteur d'activité. Globalement, quel que soit la catégorie, nous faisons les conclusions suivantes :

— La population masculine est plus à risque que la population féminine,
— Les prêts *Delega* sont plus à risque que les prêts CQS.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon alternance et qui m'ont aidé lors de la rédaction de ce mémoire.

Je voudrais dans un premier temps remercier, mon directeur de mémoire Monsieur Olivier Lopez, directeur de l'Institut de Statistique de l'Université de Paris, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je remercie également toute l'équipe pédagogique de l'Institut de Statistique de l'Université de Paris et les intervenants professionnels responsables de ma formation, pour avoir assuré la partie théorique de celle-ci.

Je tiens à témoigner toute ma reconnaissance à mon tuteur en entreprise. Monsieur Ismail Elhadni, m'a beaucoup appris cette année. Il a partagé avec moi ses connaissances et expériences dans le milieu de l'actuariat, tout en m'accordant sa confiance et une large indépendance dans l'exécution de missions valorisantes.

Je voudrais aussi exprimer ma reconnaissance envers mes collègues du service Actuariat International de Sogecap qui m'ont apporté leur soutien intellectuel tout au long de ma démarche.

Je remercie aussi mes parents pour leur soutien constant et leurs encouragements.

Table des matières

Introduction

En actuariat, l'étude, le calcul et la compréhension des taux d'incidence de la sinistralité sont primordiaux. En effet, ceux-ci sont nécessaires, pour chaque nouveau produit, afin de déterminer par exemple, la tarification adéquate et le provisionnement nécessaire. Régulièrement, les produits existant sur le marché doivent être étudiés, dans le but de vérifier notamment l'alignement et la cohérence de ces derniers avec del résultats prévisionnels et définitifs par ligne d'activité. Nous vérifions que ceux-ci répondent aussi aux exigences réglementaires.

Ce mémoire porte sur l'étude du taux d'incidence chômage sur le portefeuille des *Cessione del Quinto*, un type particulier de prêt italien couvert par une assurance emprunteur, crée en 1914. Dès lors, notre objectif est de mieux cerner les facteurs influençant la perte d'emploi et de prédire cette dernière de la façon la plus précise possible.

La première partie de ce mémoire présente le cadre de l'étude, notamment les caractéristiques et spécificités du produit qui en fait l'objet. L'assurance emprunteur des *Cessione del Quinto*, en forte croissance en Italie, présente des spécificités concernant les conditions de souscriptions et les garanties, que nous présenterons dans ce mémoire.

La seconde partie consiste en une analyse complète des données dont nous disposons dans le cadre de cette étude. Celle-ci est une étape très importante, permettant de nettoyer et trier notre base, de détecter d'éventuelles anomalies et d'extraire de premiers résultats et axes de réflexion.

Les troisième et quatrième partie sont consacrées au calcul des taux d'incidence de la Perte d'emploi via deux approches différentes.

La troisième partie utilise les méthodes usuelles d'analyse de survie, qui sont la méthode de Kaplan Meier et le modèle de Cox. Cette première approche, classique et essentielle, offre des premiers résultats très concluants.

Enfin, dans la dernière partie nous utilisons, pour le calcul des taux d'incidence perte d'emploi, une méthode très différente : le *Machine Learning* (ou apprentissage automatique). Il s'agit d'une technique de programmation qui utilise les probabilités statistiques afin de donner aux ordinateurs la capacité d'apprendre par eux même.

Nous étudierons les différences entre les méthodes et nous comprendrons les avantages et inconvénients de chacune d'entre elles.

Enfin, nous comparerons les divers résultats obtenus et nous conclurons notre étude.

Chapitre 1

Contexte et cadre de l'étude

1.1 Le marché italien du Cessione del Quinto

1.1.1 Le développement du marché

La naissance du produit

Le CQS (*Cessione del Quinto dello Stipendio*), est un type particulier de prêt créé en 1914, permettant à l'emprunteur d'effectuer un prêt dont le montant maximal des mensualités ne peut être supérieur au cinquième de son salaire mensuel net ou de sa rente. Il s'agit à l'origine d'un prêt spécial pour les employés de chemins de fer en Italie. Plus tard, ce produit est régulé par une nouvelle réglementation, le décret DPR n°180/50 du 5 juillet 1950, entré en vigueur le 28 juillet 1950 et constitue désormais un prêt destiné uniquement aux fonctionnaires italiens (secteur public et parapublic), dans le but de leur faciliter l'accès aux crédits.
Cette loi a été rédigée dans un contexte d'après-guerre caractérisé par une forte croissance du pays et la renaissance d'une classe moyenne, essentiellement constituée d'agents de l'Etat et des collectivités et administrations locales.

Les mensualités des Cessione del Quinto ne peuvent dépasser le cinquième du salaire. Une particularité de ce prêt est que ces mensualités sont prélevées directement à la source par l'employeur et versées tous les mois au créancier. La durée totale du prêt ne peut excéder 10ans, et n'est généralement pas inférieure à 24 mois.

Des évolutions plus tardives

En vertu de la loi promulguée par l'autorité de régulation italienne en 1993, le CQS ne doit être accordée que par des intermédiaires financiers officiels inscrits au registre légal tenu par la Banque d'Italie, anciennement le Bureau de change italien. Ainsi, seuls les établissements de crédit et les institutions de prévoyance constituées par des employés des administrations publiques, l'Institut National de l'Assurance, les compagnies d'assurance, les institutions et les sociétés et instituts de crédit à l'exclusion des sociétés de personne, des caisses d'épargne et des organismes de prêts sur gage sont autorisés à octroyer ce prêt.

En vertu du décret DPR 180/1950, le risque de défaut de l'emprunteur en cas de décès ou de

perte d'emploi, due à un licenciement ou une démission, doit obligatoirement être couvert, couverture assurée, à l'origine, uniquement par l'institut national de sécurité sociale pour les employés de l'administration public, l'INPDAP (acronyme de *Instituto Nazionale di Previdenza per i dipendenti dell'Amministrazione Pubblica*).

Rôles importants des lois de finance

D'autres modifications réglementaires ont suivi : les lois sur les finances 311/2004, 80/2005 et 266/2005. Ces textes ont notamment mis fin au monopole de l'INPDAP pour la couverture du risque d'insolvabilité de l'emprunteur, ouvrant ainsi ce marché aux assureurs privés. Par ailleurs, ces nouvelles normes autorisent à étendre l'octroi de ces prêts CQS aux employés du secteur privé ainsi qu'aux titulaires d'une rente (retraités et invalides), distinguant ainsi deux types de produits :

— CQS (*Cessione del Quinto dello Stipendio*) pour les populations actives
— CQP (*Cessione del Quinto della Pensione*) pour les retraités et titulaires d'une rente d'invalidité

Quelques années plus tard, à la suite de quelques abus quant à l'utilisation du produit, la Banque d'Italie a rappelé dans sa communication officielle de 2009 192691/09 les règles à respecter pour la commercialisation et la gestion des emprunts CQS et CQP. Entre 2015 et 2016, elle a de nouveau réglementé le registre des intermédiaires agréés en introduisant de nouvelles conditions d'admission et de contrôle supervisé.

1.1.2 Les spécificité des Cessione del Quinto

Le *Cessione del Quinto* est un produit très répandu et facile d'accès en Italie Il n'y a aucune contrainte d'utilisation de la somme prêtée, l'emprunteur peut la dépenser pour ce qu'il souhaite.

Comme déjà indiqué, les *Cessione del Quinto* représentent un moyen simple de demander un prêt personnel en Italie. En recevant l'argent directement de l'employeur ou de l'organisme de sécurité sociale, par le biais de la retenue sur salaire, les banques ne courent pas de grands risques et donc octroient aisément ce prêt, sans exiger de garanties supplémentaires. Ainsi, le CQS est un produit très régulé par la loi notamment en termes d'acceptation de risques, des âges limites, d'assurances *etc.*...

Plusieurs acteurs pour un produit

Les acteurs de ce produit sont au nombre de quatre :

— L'emprunteur : un retraité (dans le cas d'un CQP) ou employé (dans le cas d'un CQS) du gouvernement, du secteur public, des administrations publiques, du secteur parapublic ou du secteur privé. Il verse une part allant jusqu'au cinquième de son salaire ou de sa retraite à l'institution financière prêteuse.
— L'employeur (disposant d'une notation de 1 à 6) ou l'institution de retraite et de prévoyance : D'après les lois régulant le produit, il est dans l'obligation, d'accepter la demande de transfert, par l'employé (ou le retraité), du cinquième de son salaire (ou de la rente) vers la Banque qui fournit le prêt. Cette obligation persiste tout au long du plan d'amortissement tant que la paie est suffisante.
— La banque ou l'Institution financière : Elle fournit le prêt personnel, d'un montant maximum de 80 000 euros à taux fixe pour une durée allant de 24 à 120 mois.

— L'assurance : elle est Obligatoire et couvre le risque décès pour le CQP/CQS, et le risque chômage toutes causes (y compris départ volontaire de l'employé) pour le CQS. La prime est unique.

CQP	CQS
— le montant emprunté — la durée du prêt — l'âge de l'assuré	— le montant emprunté — la durée du prêt — l'âge de l'assuré — l'ancienneté de l'assuré

TABLE 1.1 – Eléments intervenant dans le calcul de la prime CQ

Le fonctionnement du Cessione del Quinto

Dans le cas des salariés, l'assurance sert à couvrir les banques contre les éventuelles pertes d'emploi du demandeur, auquel cas la banque serait indemnisée du montant résiduel du prêt. La subrogation permet à l'assureur qui a payé l'indemnisation de se retourner contre le responsable du dommage.

Ainsi, en cas de perte d'emploi dûe à un licenciement abusif, sans raison valable ou en cas de faillite, après avoir indemnisé la banque ayant octroyé le prêt, l'assurance peut se retourner contre l'employeur de l'emprunteur. S'il n'est pas en mesure de rembourser le prêt, il sera déclaré comme mauvais payeur et exclu du circuit du crédit. En effet, pour les entreprises privées et parapubliques, l'assureur dispose d'un système d'évaluation interne basé sur les états financiers, les données du marché, les conventions collectives et le nombre d'employés des sociétés. Alors, un système de notation est mis en place, permettant à l'assureur de refuser la couverture d'assurance dans le cas d'entités, fonds de pension ou autres institutions évaluées comme présentant des risques élevés. Ceux-ci doivent être indiquées, dans une liste noire à mettre à jour si nécessaire. Ce système permet d'atténuer les risques dus à l'insolvabilité ou la faillite de l'employeur.

Dans le cas d'un licenciement pour un motif justifié, la compagnie d'assurance prélève un montant appelé TFR (*Trattamento di Fine Rapporto*) et peut réclamer un remboursement complémentaire en cas de nouvel emploi entamé rapidement. Nous détaillons ce point dans la partie suivante. Dans le cas des retraités, le type de police requis est différent. Le montant de la police d'assurance varie considérablement selon l'âge.

Deux particularités importantes : le Delega et le TFR

Le fond TFR

Pour les employés du secteur privé et para-publique uniquement, le TFR ou *Trattamento di Fine Rapporto* est l'indemnité de départ, régie par le code du travail italien . Lorsqu'une rupture du contrat de travail est à l'initiative de l'employeur ou non, ce dernier a l'obligation de verser une indemnité de départ au salarié. Celle-ci comprend entre autres les congés annuels accumulés non utilisés, le paiement de la période de préavis obligatoire, une aide à la recherche d'un nouvel emploi... Le TFR correspond à environ 1 mois de revenu brut par année d'ancienneté. Il alimente un compte "épargne" dont l'assuré pourra bénéficier lorsqu'il quittera l'entreprise.

Le montant d'un prêt *Cessione del Quinto* est déterminé via le TFR. L'employé n'est pas autorisé à accéder à son TFR avant la fin du prêt CQS.
Ce montant viendra en déduction du Capital Sous Risque à payer par l'assureur lors d'un éventuel sinistre perte d'emploi, et ce, grâce au nantissement du TFR dont bénéficie l'assureur. Nous obtiendrons donc le montant suivant :

Capital sous risque de l'assureur pour le risque chômage = Capital restant dû – TFR

Si l'entreprise fait faillite, le TFR ne peut plus être utilisé.

Comme indiqué précédemment, si l'emprunteur réalise à la fois un prêt CQS et un prêt *Delega*, en cas d'insolvabilité le TFR est consacré prioritairement au remboursement du CQS.

Le prêt complémentaire Delega

Dans le cas du CQS uniquement, il existe un prêt complémentaire, appelé *Delega*, qui consiste en un second prêt additionnel au CQS ne présentant pas les mêmes caractéristiques.
Nous notons notamment qu'il requiert une autorisation par l'employeur (non requise pour CQS). Le montant des mensualités n'est pas régulé par la loi (alors qu'il l'est pour le CQS). En cas de complément à un prêt CQS, le TFR est consacré prioritairement au remboursement du CQS

Petit comparatif avec les prêts personnels classiques

Les produits CQP/CQS diffèrent beaucoup des prêts personnels classiques. Ainsi nous pouvons constater les éléments suivants.

	Prêts classiques	**CQ**
Montant accordé	limité à 30 000 euros	limité à 55 000 euros ou 100 000 euros si *Delega*
Versements	effectués par l'emprunteur	effectués automatiquement par l'employeur
Assurance	Optionelle	Obligatoire
Evaluation de risque de crédit pour l'emprunteur	Obligatoire	Optionelle
Evaluation de risque de crédit pour l'employeur	Optionelle	Obligatoire
Risque de défaut volontaire du débiteur	Très élevé	Nul

TABLE 1.2 – Comparaison des prêts personnels classiques et du produit CQS/CQP

Les autres spécificités liées à ce produit

Si l'emprunteur est en contrat à durée déterminée, l'ensemble des versements doivent être effectués avant la fin du contrat de travail. Tout cela sans dépasser des mensualités représentant un cinquième du salaire. Dans certains cas, il peut être utile d'avoir un garant pour couvrir le prêt, ce qui permet de dépasser la limite du contrat, ou de l'obtenir plus facilement.

Le respect de la règle des"deux cinquièmes" est essentiel quant à l'utilisation des *Cessione del Quinto* : il n'est possible de racheter un prêt *Cessione del Quinto* que si la durée de vie écoulée du prêt est supérieure ou égale aux deux cinquièmes de la durée de vie totale prévue initialement (Loi 180/50 du code italien).
Les CQ constituent un secteur dynamique et en croissance. Une loi, dès lors spécifique à l'Italie, pourrait être exportée comme modèle vers les autres pays afin d'y réduire le coût des crédits à la consommation. L'association bancaire d'Italie ABI et l'ASSOFIN (crédit à la consommation) ont développé un dossier détaillé soumis au comité de Bâle démontrant que la moyenne de l'indice de risque de crédit diminuerait de cinq fois en utilisant la formule de CQP/CQS. Ainsi, le CQS pourrait bientôt être exporté hors Italie présentant ainsi un avantage pour les consommateurs (les taux sont plus bas) et comporte moins de risque pour le système bancaire.

Différentes hypothèses sont en cours d'évaluation afin de mettre à jour le produit. La récente crise économique a remarquablement encouragé cette forme de prêt qui est actuellement fournie non seulement par des organismes spécialisés mais aussi par presque toutes les banques. Ce projet sera soumis au Parlement italien, mais les discussions sont toujours en cours. Les points clés de la proposition de réforme sont les suivant :

— l'augmentation du montant des mensualités de 20% à 30%
— la supression de la *Delega*
— la supression de la taxe sur les garanties (y compris les couvertures d'assurance)

1.1.3 Les contrats d'assurance des Cessione del Quinto

Les *Cessione du Quinto* sont protégés, comme le prévoit la loi, par un programme d'assurance qui garantit le financement du risque de crédit lorsqu'il n'est pas possible de poursuivre les paiements en raison du décès du titulaire et/ou de la perte d'emploi pendant la durée de couverture d'assurance.

Le client, lors de la proposition contractuelle, fournit des déclarations sur son état de santé, ainsi que, en cas de sinistre, toutes les informations et la documentation nécessaires pour permettre l'activation des couvertures d'assurance en compensation du crédit résiduel.

Nous nous concentrons désormais, dans la suite de ce mémoire, sur les CQS, destinés aux actifs, le produit CQP étant réservé aux retraités.

Ci-dessous les caractéristiques de souscription aux contrats d'assurance et la politique de sinistre attachée à ce dernier :

	CQS
Assuré	L'emprunteur
Parties prenantes	L'emprunteur et L'employeur
Durée du prêt	24 à 120 mois
Sélection à l'entrée	— Questionnaire médical simplifié et/ou questionnaire médical complet — ancienneté de 12 mois
Montants assurés	— Montant maximum assuré sur le CQS : 55 000€ — Montant maximum assuré sur le Delega : 45 000€ — En cas de décès : le capital restant dû le jour du sinistre — En cas de perte d'emploi : le capital restant dû le jour du sinistre diminué du montant du TFR de l'emprunteur (pour le secteur privé)
Prime	Prime unique payable d'avance. Calculée sur la base du montant emprunté, de l'âge et de l'ancienneté
Gestion des sinistres	En cas de perte d'emploi, l'institution financière se doit de prélever le TFR auprès de l'employeur. Si ce dernier n'est pas suffisant, l'assureur règle le reliquat.

TABLE 1.3 – Principales caractéristiques des CQS

Le questionnaire médical concernant l'état de santé de l'emprunteur CQS constitue une étape majeure pour l'assureur dans le calcul des cotisations et des garanties. Les questions portent sur l'état de santé passé et actuel, notamment les traitements médicaux et pathologies récentes.

1.2 Les contrats Cessione del Quinto commercialisés par SOGECAP/SOGESSUR

1.2.1 Place de SGI Italy sur le marché Cessione del Quinto

La Société Générale, l'un des groupes bancaires les plus importants, est présente dans plus de 67 pays à travers le monde. A ce jour, le groupe compte plus de 31 millions de clients.
Depuis plus de dix ans, la *Société générale Insurance* ou SGI apporte son savoir-faire dans le domaine de la bancassurance internationale, disposant de l'expertise et de la solidité d'une des plus grandes banques sur le marché français.

Société Générale Insurance est présente en Italie depuis 2010 via sa succursale située à Milan. Elle propose une large gamme de produits et de services d'assurance sur les activités Vie et Non Vie, dans

le but de répondre aux besoins du marché italien.

Le marché global est sur une trajectoire croissante depuis 2015. Le secteur a prouvé que de plus en plus d'investisseurs institutionnels étrangers et attractifs regardent les opérateurs Italien avec beaucoup d'intérêt. Le marché du CQS est un secteur dynamique qui vise à croître et à jouer un rôle de premier plan dans le secteur du crédit.
Le montant des crédits accordés sur le marché CQ est en augmentation constante. En 2019, presque 5 500 M d'euros de crédit CQ sont accordés, comprenant 48% pour les CQP et 52% pour les CQS.

Société Générale Insurance en Italie a démarré la commercialisation des produits Cessione del Quinto en 2013. Son premier partenariat fût avec une société spécialisée dans le crédit à la consommation, que nous nommerons dans ce mémoire, par soucis de confidentialité, Partenaire1. La convention signée concerne uniquement le produit CQP. Puis, elle signe son premier partenariat externe avec un second partenaire, d'ores et déjà positionné dans le marché italien CQP. En 2015, la SGI débute la commercialisation des CQS.

Les années 2016 et 2017 permettront l'élaboration de nouveaux partenariats externes pour le CQP et le CQS respectivement.
La part de marché de SGI Italie est en forte croissance.

L'activité CQP/CQS représente une part importante du business de SGI Italie et reste un levier de croissance majeur sur l'horizon budgétaire via le développement de l'activité elle-même et via la contractualisation avec de nouveaux partenaires distributeurs pour étendre, par la suite, la gamme produit.

1.2.2 Etude des contrats d'assurance des Cessione del quinto par SGI

Les conditions de la couverture d'assurance

La couverture d'assurance est fournie par l'assureur à l'assuré dans le cadre du prêt, à condition que l'emprunteur ait souscrit à un prêt CQS dont la durée est comprise entre 24 et 120 mois, qui prévoit des remboursements mensuels et dont le montant ne dépasse pas les montants maximums. Il doit fournir au preneur d'assurance de nombreux documents attestant de son identité, sa situation professionnelle et ses revenus. S'il n'est pas fonctionnaire, il doit aussi indiquer le montant de son TFR.

Ci-dessous le détail des conditions de la couverture d'assurance :

Age maximum de l'emprunteur à l'expiration du prêt	— Secteur public / CQS : 72 ans — Secteur public /Delega : 66 ans — Secteurs Parapublics et privés / CQS et Delega : 66 ans — Pour la marine marchande et les forces armées : 60 ans
Ancienneté minimale	6 mois pour le CQS et 12mois pour le Delega (la période d'essai doit être terminée)
Santé	Questionnaire médical simplifié et/ou questionnaire médical complet selon le montant emprunté et l'âge
Montant du salaire mensuel net minimum	Le salaire net de retenues doit être $\geq$ 500€ (= seuil de survie)
Exigences de travail	L'assuré ne doit pas : — être en contrat de formation, d'essai, d'apprentissage ou de stage — être un partenaire de travail (sauf partenaires de travail disposant d'un revenu régulier certifié) — être en congé ou en suspension d'emploi — faire l'objet de mesures disciplinaires — être en congé de maternité De plus : — dans le cadre d'un contrat de travail temporaire le remboursement du prêt doit être effectué avant la cessation d'emploi — en tant que participant à un Fonds de pension auquel son indemnité de départ ou une partie de celle-ci est transférée, la partie cédante doit signer une déclaration selon laquelle, en cas de cessation d'emploi , elle accepte que le Fonds de pension transfère une partie des sommes accumulées à la partie contractante, pour rembourser le prêt résiduel.
Pourcentage de la retenue maximale sur le salaire	20% du salaire pour le CQS et 40% si le CQS est complété par un *Delega*. Seuls les prélèvements obligatoires sont pris en compte (d'éventuels versements volontaires, ne sont pas inclus)
Limite absolue du montant brut	pour le CQS, la limite est de 55 000€ et pour le Delega elle est de 45 000€
Montant brut maximum accordé	Pour les employés du secteur public, il vaut la limite absolue (soit 55 000 euros pour le CQS et 45 000 pour le Delega). Pour les autres employés, il est calculé en fonction du TFR et du rating de l'employeur

TABLE 1.4 – Critères de souscription Cessione del Quinto

La pratique du marché veut que le montant maximum de l'emprunt correspond à : min(limite absolue; [Rating d
Montant du TFR])

La couverture d'assurance prend effet 24 heures après la date d'octroi. Sa durée ne doit pas dépasser la durée du prêt. Elle reste effective pendant toute la durée du plan de remboursement du prêt à condition que la prime soit payée et prend fin une fois le prêt remboursé ou bien en cas de survenance d'un sinistre (décès ou perte d'emploi).

Les exclusions et limitations du contrat sont les suivantes :

— Le non-paiement ou le retard de paiement des échéances par l'employeur
— La non-validité ou l'inexistence du prêt
— Le défaut d'octroi du montant du prêt
— La falsification documents
— Le non-respect (en tout ou en partie) de la procédure d'activation de la couverture d'assurance. En effet, le consentement de l'assureur pour l'octroi de la couverture d'assurance est conditionné à la véracité des déclarations faites par le cédant/déléguant.

Le montant maximum payable par l'assureur est égal à 100 000 € (pour CQS et *Delega*). Les intérêts de retard éventuels étant exclus des prestations payables.

Les garanties rattachées au produit

Plusieurs types de garanties sont offertes selon les cas :

En cas de rachat du prêt par l'emprunteur :

En cas de remboursement anticipé ou de transfert du prêt, la couverture d'assurance cessera d'être effective et l'assureur remboursera une partie de la prime liée proportionnellement à la période d'assurance non prise en charge.
La partie à rembourser est égale à : $\pi \times \frac{Dette}{MI} + C$ avec :

— π = prime de risque correspondant à la période d'assurance résiduelle par rapport à l'échéance initiale
— Dette = dette résiduelle à la date de l'extinction anticipée
— MI= Montant initial du financement
— C= coûts correspondant à la période d'assurance résiduelle par rapport à l'échéance initiale

En cas de décès de l'emprunteur :

En cas de décès, il est obligatoire de fournir un certificat de décès et des documents pour 40

l'organisme de crédit. Alors, l'assureur verse à la société octroyant le prêt le montant du capital restant dû à la date du décès, calculé à partir du montant du prêt, y compris les intérêts et les frais. Le taux annuel nominal (TAN) est mis en place par l'organisme de crédit au moment du prêt, pour permettre de calculer les intérêts.

En cas de perte définitive d'emploi :

En cas de perte définitive d'emploi, avant de solliciter l'assureur, l'emprunteur doit respecter une série d'obligations visant à recouvrer sa dette. Il doit aussi transmettre les documents nécessaires pour

vérifier l'existence et le montant actuel du remboursement résiduel.

L'assurance demande au client de payer tout ou en partie la dette restante via des actions de recouvrement. Il exige notamment le TFR auprès de l'employeur (pour le secteur privé) et le déplacement du prêt lors du départ à la retraite (pour le secteur public). Ainsi, l'employeur doit retenir tout le montant accumulé par l'employé de l'entreprise et payer cette somme à la banque (il se réserve le droit d'utiliser la dernière paie, le solde tout compte, les congés non payés...).

Dans le cas où un emprunteur a un prêt CQS ainsi qu'un prêt Delega, les sommes disponibles pour le recouvrement, notamment le TFR sont destinées en priorité à recouvrir le prêt CQS. De ce fait, comme nous le verrons plus en détail dans les parties suivantes, le risque relatif au Delega est sensiblement plus élevé que celui des CQS.

Une fois ces étapes effectuées, en cas d'impossibilité de paiement avérée, l'assureur s'engage à procéder à l'indemnisation dans les 37 jours suivant la réception de la documentation complète, délai permettant de laisser le temps aux actions de recouvrement citées.

Les CQS sont des prêts présentant des caractéristiques particulières. Ils sont prélevés directement sur salaire et les prélèvements ne peuvent excéder le cinquième de celui-ci. Les produits CQ sont obligatoirement accompagné d'une assurance. Celle-ci couvre les risques d'insolvabilité de l'emprunteur, en raison d'un décès où d'une perte d'emploi.
Nous nous intéressons, dans la suite de ce mémoire à la couverture de la Perte d'emploi.

Chapitre 2

Données disponibles et statistiques descriptives

Dans la première partie, nous avons présenté les généralités sur le produit *Cessione del Quinto* en Italie commercialisé par SGI Italie, ainsi que les assurances obligatoires associées et leurs caractéristiques.

Dans cette section, nous nous concentrerons essentiellement sur les données d'assurance mises à notre disposition pour la réalisation de ce mémoire.

Notre démarche statistique se fera en trois étapes :
- Préparation de la base d'étude
- Traitement des variables : analyse univariée, analyse bivariée et analyse multivariée
- La modélisation : via les méthodes classiques puis le Machine Learning

Il est nécessaire de préparer nos données et d'effectuer des tests quant à leur fiabilité car, étant donné le volume important de données, nous détectons souvent des incohérences. Dans ce cas, il faut les corriger afin de minimiser le biais.
Nous choisissons alors, selon la situation, de corriger les données par des approximations à l'aide d'autres variables, par des calculs de moyennes, par des suppressions *etc* ...

2.1 Description des bases de données (des assurés et des sinistres)

2.1.1 Données d'étude reçues et utilisées

Les bases de données des assurés et celles des sinistrés sont disponibles. Elles contiennent l'historique de tous les contrats d'assurance *Cesionne del Quinto* commercialisés par SGI depuis août 2015 jusqu'en début décembre 2019. Ces bases se sont avérées assez complètes ; en effet, elles contiennent beaucoup de paramètres (soixante-quinze).
Par soucis de simplification, uniquement les variables qui sont jugées plus pertinentes pour notre étude, sont présentées.

La base de données des assurés : Elle présente 10 087 contrats caractérisés par leur numéro de police ainsi que de nombreuses informations. Elles présentent :

- Des caractéristiques concernant l'employé, l'emprunteur du prêt Cessione Del Quinto :
  - âge à la souscription
  - date de naissance
  - sexe

  L'âge à la souscription et le sexe sont conservés.
  - La catégorie socio-professionnelle (publique, parapublique ou privée). Cette variable sera, comme nous le verrons dans la suite, primordiale.
  - Pour les contrats concernant des salariés du secteur privé et parapublique : la valeur du TFR à la souscription du contrat
- Des caractéristiques concernant l'employeur :
  - description de la compagnie et des détails sur celle-ci, qui ne sont pas nécessairement utile et ne sont donc pas conservés
  - Son coefficient multiplicateur (ou *rating*)
- Des caractéristiques concernant le prêt :
  - le montant assuré ainsi que le montant brut à rembourser (le montant assuré majoré des intérêts). Les deux informations étant entièrement corrélées et ainsi redondantes, seul le montant assuré est conservé.
  - la durée initiale prévue
  - le montant des mensualités de remboursement correspondantes. Il est simplement le résultat de la division du montant brut à rembourser par la durée du contrat *Cessione del Quinto*, nous ne conserverons pas cette variable.
  - le type de prêt : il peut s'agir d'un CQS, ou bien d'un prêt secondaire, un *Delega*, comme présentés dans le chapitre précédent.
  - Le nom du partenaire, c'est-à-dire l'institution financière ayant octroyé le prêt. Il y en a cinq différents, avec une prépondérance du Partenaire1, qui correspond à plus de 80% des contrats. De ce fait, nous conserverons cette variable avec seulement deux modalité : Partenaire1 ou *Others* (la catégorie *Others* regroupant l'ensemble des autres partenaires). C'est un choix tout à fait cohérent, notamment car la même distinction est faite par Sogecap pour la réalisation des documents comptables et financiers.
- Des caractéristiques du contrat d'assurance :
  - Le montant de la prime versée à l'assureur (brute et nette des taxes). Cette prime est unique, versée à la souscription. Après réflexion, nous ne conservons pas cette variable comme variable explicative mais elle nous sera utile dans le cadre de la vérification de nos données et notamment la réconciliation comptable.
  - La date de souscription : Comme expliqué précédemment, le produit étudié est très récent, nous disposons donc d'un historique assez faible, débutant en 2015. La base est extraite en début décembre 2019.
  - Lorsque les contrats sont déjà terminés, nous disposons de leur date de clôture. Les contrats clôturés peuvent être de trois types :
    - arrivés à leur terme
    - dus à un remboursement anticipé
    - dus à la survenance d'un sinistre qui a mis un terme au contrat d'assurance

    ,

La base de données des sinistres : Elle compte 204 enregistrements. Les contrats sont toujours caractérisés par leur numéro de police, qui nous permet de retrouver toutes les informations concernant

le contrat (le partenaire financier, le produit assuré...). Nous disposons des informations suivantes :

- la date de survenance et la date de déclaration du sinistre. Nous rappelons que la base est extraite en début décembre 2019, aucun sinistre ni déclaration n'a donc lieu après. Par ailleurs, bien que la commercialisation du produit ait débuté en 2015, le premier sinistre a eu lieu en 2016.
- Le statut du sinistre : en cours, accepté ou refusé. Finalement, dans notre base de données, tous les sinistres sont soit en cours soit acceptés, il n'y a pas de sinistre refusé. Il sera très important de considérer aussi les sinistres en cours, afin d'éviter de biaiser nos résultats.
- le capital restant dû au moment de chaque sinistre. Il est obtenu en fonction du montant brut à rembourser par l'emprunteur du Cessione del Quinto et de la durée écoulée depuis le début du contrat.

Point d'attention sur le déroulement de l'étude : Les garanties mises en jeu dans cette base sont de deux types : décès et perte d'emploi. L'objet de ce mémoire étant l'étude du taux d'incidence de la perte d'emploi sur notre feuille, nous ne conservons que les enregistrements mettant en jeu la garantie perte d'emploi. Ainsi, nous nous restreignons aux 177 enregistrements concernant un sinistre perte définitive d'emploi.

Nous souhaitons étudier le taux d'incidence perte d'emploi pour notre portefeuille. Ainsi, nous regroupons nos deux bases de données en transformant la variable Sinistre une variable à deux modalités, de valeur "Non" si aucun sinistre lié au chômage n'a eu lieu pour ce contrat et de valeur "Oui" si un sinistre a eu lieu.

Nous calculons aussi la duration de chaque contrat, qui nous sera utile pour la suite. Il s'agit de la durée écoulée depuis le début du contrat. Selon les contrats, celle-ci peut correspondre à la différence entre la date de souscription de contrat et sa date de clôture, ou bien, si le contrat est encore en cours, cette duration correspond à la durée écoulée depuis la souscription du contrat jusqu'à la date d'extraction de la base, soit fin décembre 2019. Elle est calculée en mois. Cette duration permet de connaître la durée actuellement écoulée pour chaque contrat sur laquelle nous nous basons pour notre étude.

Pour la suite, toutes les périodicités considérées seront mensuelles.

2.1.2 Les tests de cohérence

Vérification de la complétude et de la conformité des données

Nous vérifions l'adéquation entre les formats de variables de nos bases de données et ceux requis. Les dates, les montants et autres modalités des variables de notre base de données sont conformes, aucune valeur n'est inappropriée.

Nous effectuons des statistiques donnant la fréquence de valeurs manquantes pour chacune de nos variables. Nous disposons finalement de données très complètes, à l'exception des informations concernant le TFR et son coefficient multiplicateur. Naturellement, seules les personnes travaillant pour des sociétés privées ou para-publiques disposent d'un TFR. Pour 516 individus de cette catégorie nous ne

disposons pas du coefficient multiplicateur du TFR et pour 400 d'entre eux nous ne disposons pas non plus du TFR lui-même.

Cohérence des données numériques dans le contexte des Cessione del Quinto

L'étude de la cohérence des données permet de vérifier deux aspects : la cohérence logique et numérique des données, mais aussi la cohérence dans le cadre du produit que nous étudions. Nous effectuons une série de tests sous Excel permettant d'identifier puis de corriger, de la manière la plus appropriée, les incohérences.

La base de données des assurés

Nous vérifions la cohérence des dates :

- La date de début du prêt est supérieure à août 2015 (date de début de commercialisation des contrats) et inférieure à décembre 2019
- La date de début du prêt est inférieure à la date de fin du prêt
- L'emprunteur est majeur à la date de début du prêt

Dans le contexte des Cessione del Quinto, nous vérifions les critères sur l'âge de l'emprunteur à la fin du prêt. Pour les employés du secteur publique l'âge à la fin du prêt doit être inférieur ou égal à 72 ans pour les CQS et à 66 ans pour les *Delega*. Concernant les employés du parapublic et privé ayant un prêt CQS ou *Delega*, l'âge maximum à la fin du prêt doit être de 66 ans pour les hommes et 63 ans pour les femmes.

Nous avons repéré moins de cinq contrats dont les emprunteurs ont plus que l'âge requis à la souscription et ce d'à peine un à deux ans. Ainsi, nous n'avons pas effectué de correction, considérant qu'il s'agit de simples erreurs de procédure de souscription, et que les enregistrements en question sont alors à prendre en compte dans le risque.

Nous vérifions que la durée du contrat est comprise entre 24 et 120 mois, un critère obligatoire pour l'octroi d'un prêt Cessione del Quinto.

Les coefficients multiplicateurs caractérisent le *rating* de la société. Celui-ci dépend du classement et du type d'entreprise considérée. Nous vérifions qu'il s'agit d'un nombre entier compris entre 1 et 6.

Nous vérifions la cohérence des montants :
Le montant brut doit correspondre au montant assuré majoré des intérêts, en fonction des taux en vigueur. Nous vérifions que :

- Le Montant brut est inférieur au montant maximum absolu
- Le Montant brut est inférieur au montant autorisé au vu de la société employeur

  - Pour les employés du secteur public ce montant est le même que le montant maximum absolu.
  - Pour les autres, en cas de CQS seulement ou de Delega complémentaire, le TFR et coefficient multiplicateur interviennent. Le montant total maximum autorisé (CQS + Delega) est le minimum entre le montant maximum total absolu (CQS + Delega) et la valeur du TFR multiplié par son coefficient multiplicateur.

  Naturellement, pour les 516 enregistrement pour lesquels nous ne disposons pas du coefficient

multiplicateur du TFR, il ne nous est pas possible de calculer le montant maximum autorisé, nous retiendront donc le maximum absolu.

Le montant remboursé mensuellement multiplié par la durée du prêt doit être égal au montant brut.

La base de données des sinistres

En fonction du numéro de contrat associé à chacun des sinistres nous vérifions la cohérence avec la base de données des assurés : informations sur l'emprunteur, nom du partenaire, type de produit...

Nous vérifions la cohérence des dates :

- La date d'occurrence du sinistre doit être postérieure à la date de début de contrat
- La date de déclaration du sinistre doit être postérieure à la date d'occurrence du sinistre
- La date de déclaration du sinistre doit être inférieure à la date de fin du prêt
- La date de clôture du contrat correspond à celle de déclaration du sinistre. Elle ne peut être antérieure sinon l'assuré ne serait plus couvert au moment du sinistre, ni postérieure car la survenance de l'événement chômage de l'employé entraîne, après indemnisation, la clôture du contrat d'assurance.

Nous notons une incohérence pour seulement 1 sinistre, dont la date de déclaration du sinistre est antérieure à la date d'occurrence. Au vu du produit que nous étudions, il est tout à fait envisageable que la date de chômage soit connue antérieurement à sa réalisation. Nous envisageons en effet un possible préavis : période de prévenance que l'employeur doit respecter avant de rompre le contrat de travail. Nous conservons donc les informations enregistrées telles quelles.

Nous vérifions la cohérence des montants : Les capitaux restants dus au moment de chaque sinistre correspondent au montant initial emprunté auquel est déduit l'ensemble des échéances déjà réglés.

La réconciliation comptable

La réconciliation comptable consiste à rapprocher deux comptes. Elle permet de comparer les montants totaux notifiés dans les ressources comptables aux valeurs réellement présentes dans notre base de données. Elle corrèle les données des deux comptes et met les différences en évidence, nous donnant une idée assez générale des montants totaux et de leur cohérence.

Nous comparons les données présentes dans notre base, tant en ce qui concerne les primes et les sinistres, aux valeurs présentes dans les fichiers comptables.
Nous obtenons des résultats très cohérents, avec de faibles écarts. Un écart plus important est calculé en 2019, celui-ci est néanmoins tout à fait logique : les fichiers comptables sur lesquels nous avons travaillé sont réalisés à la fin de chaque année tandis que notre base de données se termine en début décembre. Ainsi, celle-ci ne comprends pas l'ensemble des primes gagnées en décembre 2019. Il en va de même pour les sinistres, en plus des sinistres réalisés non encore déclarés, nous ne disposons pas des sinistres réalisés en décembre 2019.

	Montant des primes	**Montant des sinistres**
2015	-	-
2016	1	-
2017	2	-
2018	5	1
2019	14	3

TABLE 2.1 – Variations observées lors de la réconciliation comptable (en%)

2.2 Statistiques descriptives

Nous disposons désormais de notre nouvelle base de données, regroupant celle des assurés et des sinistrés. Les valeurs aberrantes et incohérences numériques ont été traitées. Les variables explicatives que nous utilisons sont :

— le sexe de l'emprunteur
— sa catégorie socioprofessionnelle
— son âge à la souscription
— la date de souscription
— le montant emprunté
— la durée de l'emprunt
— le type d'emprunt

La variable que nous cherchons à modéliser est la sinistralité.

Nous avons, dans notre base de données, un mélange de variables qualitatives et quantitatives.

L'âge, les primes versées, le montant assuré, la date de souscription, la durée du contrat, le coefficient multiplicateur du TFR, le TFR lui-même sont numériques, sous la forme de facteurs ou sous la forme d'un nombre entier.
Les autres variables sont, quant à elles, des variables qualitatives ordonnées.

Nous supprimons la variable "Numéro de contrat". Elle a beau être identifiée comme numérique, elle est en réalité qualitative et n'apporte ici aucun renseignement, si ce n'est le nombre d'observations total, que nous connaissons.

Que ce soit pour les variables qualitatives ou numériques, nous commençons avec les statistiques descriptives, qui donnent les effectifs des variables qualitatives et, pour les variables numériques, des indicateurs de centralité et de dispersion comme la moyenne et les quartiles.

Nous effectuons les statistiques descriptives sur la base de données des assurés et celle des sinistres.

2.2.1 Analyse univariée – La population des assurés

La souscription des prêts

La souscription des prêts *Cessione del Quinto* débute au troisième trimestre de 2015, en effet le lancement du produit a lieu en août 2015. Naturellement, la production cette année-là est assez faible par rapport aux quatres générations qui suivent. Les années de 2016 à 2019 étant quasiment complètes

(notre base de données n'inclut pas le mois de décembre 2019), nous appuierons sur cette période nos raisonnement permettant d'expliquer les comportements des emprunteurs pour leurs quatre premières années de prêt *Cessione del Quinto*.

Depuis 2016 le nombre de prêts souscrits est en augmentation permanente, avec une hausse très importante en 2019.

2015	298
2016	2022
2017	2072
2018	2275
2019	3420

TABLE 2.2 – Nombre de prêts souscrits par année

Nous étudions la souscription des prêts *Cessione del Quinto* au fil des trimestres, afin de détecter une éventuelle saisonnalité.

Nous constatons une similarité entre les années 2016 et 2018. Les nombres de contrats sont très proches et la tendance est la même durant l'année, à savoir une légère augmentation des souscriptions aux deuxièmes et quatrièmes trimestres.
L'année 2017 compte, en moyenne, un nombre de contrats souscrits presque similaire aux années précédemment évoquées mais la tendance ressemble d'avantage à celle de 2019 : un nombre de contrat plus importants aux deuxièmes et troisièmes trimestres.
Nous notons en 2019, une très forte augmentation de la souscription durant toute l'année.Chaque année, beaucoup de souscriptions ont lieu pendant le deuxième trimestre.

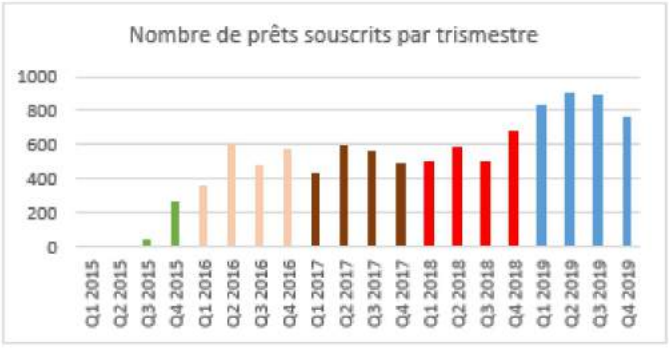


FIGURE 2.1 – Nombre de prêts CQS souscrits par trismestre

Les catégories Socio-professionnelles

Nous observons la répartition de la base de données par catégorie socio-professionnelle des employés auxquels le prêt *Cessione del Quinto* a été octroyé. Les employés du secteur parapublic sont très peu nombreux, ils représentent 5% de la population, tandis que les employés du secteur privé, près de 45%, et employés de la fonction publique , plus de 50% de la population totale, sont bien plus nombreux.
Le portefeuille est composé majoritairement d'employés du secteur public, suivis de près par les employés du secteur privés.

FIGURE 2.2 – Répartition de la catégorie Socioprofessionelle des assurés

L'âge à la souscription

La base de données étudiée présente des assurés d'âge entre 21 ans et 68 ans. Nous remarquons que nous avons peu d'observations pour les jeunes âgés de 25 ans et moins, et pour les personnes âgées, de plus de 61 ans.
L'âge moyen est de 48 ans.

count	10 087
Mean	48
Std	9
min	21
25%	42
50%	49
75%	55
Max	68

TABLE 2.3 – Statistiques descrptives de l'âge à la souscription

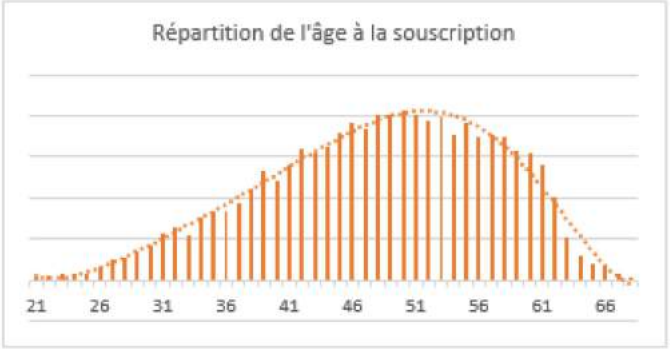


FIGURE 2.3 – Répartition de l'âge à la souscription

Le montant assuré

Les montants assurés présentent une disparité importante : ceux-ci s'étendent de 1 890€ à 84 926€. La moyenne est basse, autour de 19 000€. Ainsi les prêts sont concentrés en dessous de 40 000€, très peu sont supérieurs à ce montant.

count	10 087
Mean	18 907
Std	9 336
min	1 890
25%	11 567
50%	18 930
75%	24 309
Max	84 926

TABLE 2.4 – Statistiques descrptives des montants assurés

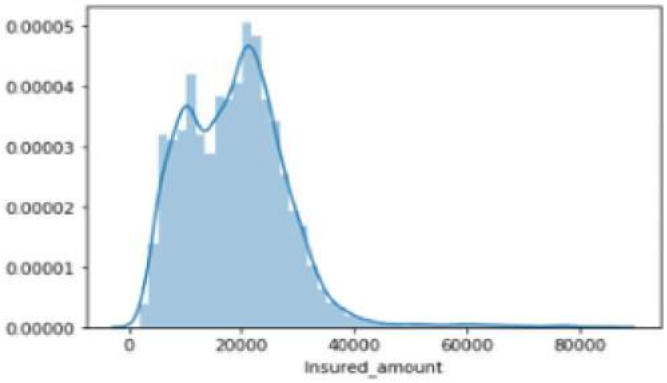


FIGURE 2.4 – Fonction de répartition des montants assurés

Nous nous intéressons aux montants des prêts selon la catégorie socio-professionnelle. Le montant moyen est plus important pour le secteur public. Il est le plus bas pour le secteur privé.

Privé	14 230
Para-public	20 238
Public	22 772

TABLE 2.5 – Montant moyen assuré par catégorie socio-professionnelle (en €)

Le sexe des assurés

Nous étudions la répartition des polices par sexe afin de s'assurer de l'uniformité de la répartition des polices. Nous remarquons que la répartition entre les deux sexes est déséquilibrée, avec plus de 60% d'hommes.

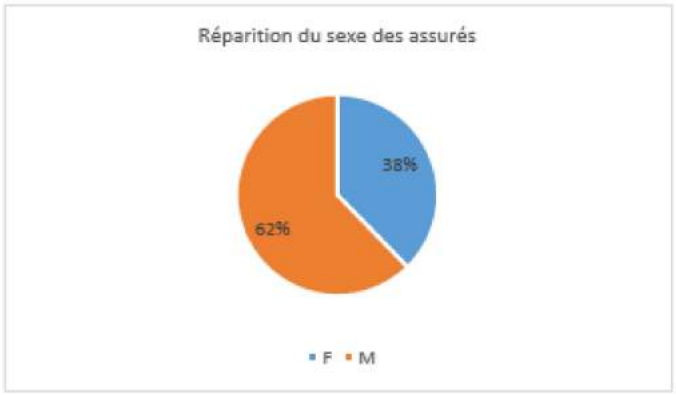


FIGURE 2.5 – Répartition du sexe des assurés

Le partenaire financier octroyant le prêt

Nous avons précédemment justifié notre choix de regrouper les partenaires financiers en deux catégories seulement, l'une dédiée au Partenaire1 et l'autre au reste des partenaires. Nous observons dans le graphique ci-dessous la prépondérance du Partenaire1, ayant octroyé plus de 80% des prêts assurés. Il est important de souligner que le Partenaire1 l'est depuis 2015 alors que les autres ne le sont que depuis 2018 voire 2019 pour certain. Ceci explique notamment l'inégalité de répartition des contrats dans le portefeuille.

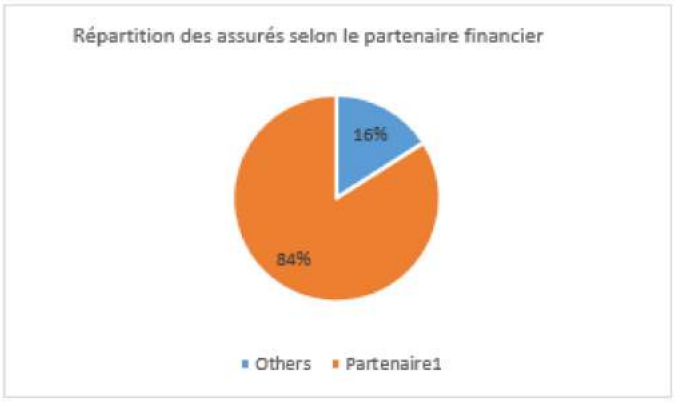


FIGURE 2.6 – Répartition des assurés selon le partenaire financier

Le type de produit

Notre portefeuille se compose de deux produits, que nous avons présenté précédemment. Ceux-ci se distinguent notamment par le fait que le *Delega* constitue un prêt secondaire. Il n'est octroyé qu'en cas de complément à un premier prêt CQS, dont le montant ne serait pas suffisant pour l'emprunteur. Ils sont donc moins courants et restreints à certaines conditions.
Le graphique ci-dessous nous indique que notre base est composée à 84% de prêts CQS. Alors nous obtenons que 68% des emprunteurs ont simplement un prêt CQS tandis que 32% ont souscrit à la fois un prêt CQS ainsi qu'un prêt *Delega*.

FIGURE 2.7 – Répartition des produits assurés

La durée des contrats

Les contrats ont une durée comprise entre 24 et 120 mois. Leur moyenne est haute, elle vaut 102.

count	10 087
Mean	102
Std	27
min	24
25%	84
50%	120
75%	120
Max	120

TABLE 2.6 – Statistiques descrptives de la durée des contrats (en mois)

Nous nous intéressons à la durée moyenne en fonction de la catégorie professionnelle. Pour le secteur privé, la moyenne est plus faible, elle est de 95 mois alors qu'elle est proche de 107 pour les secteurs publics et para-publics.

Privé	95
Para-public	108
Public	107

TABLE 2.7 – Durée moyenne des contrats (en mois)

2.2.2 Analyse univariée – Les sinistres

Répartition de la sinistralité

Comme nous l'avons déjà noté, l'une des particularités de notre base de données est que nous ne retrouvons que très peu de sinistres perte d'emploi. Nous disposons ainsi de 10 088 enregistrements dont seulement 177 ont subis un sinistre perte définitive d'emploi, soit moins de 2% de la population.

Nous sommes ainsi face à des classes déséquilibrées, un point important comme nous le verrons par la suite. C'est un point qu'il sera essentiel de prendre en compte pour la modélisation, notamment lors de la modélisation via le Machine Learning, afin de ne pas biaiser nos résultats.

FIGURE 2.8 – Répartition de la sinistralité

Nous nous intéressons aux statuts de ces sinistres. Comme le montre le graphique ci-dessous l'intégralité des sinistres clôturés ont été acceptés, il n'y a aucun refus. Seul 3% des sinistres sont en cours. Nous notons que la date d'occurrence de ces 6 sinistres est postérieure à septembre 2019, il est donc naturel que les dossiers ne soient pas encore clôturés. Nous pouvons tout de même négliger le souci du taux d'acceptation que nous considérons, au vu de nos données, de 100%.

FIGURE 2.9 – Répartition des sinistres selon leur statut

La temporalité des sinistres

Dans un premier temps, nous nous intéressons à la sinistralité observée aujourd'hui en fonction de la date de souscription des contrats.
Une grande partie des sinistres ont eu lieu sur les contrats souscrits en 2016. Les contrats souscrits arrivent ensuite sur l'année 2017, puis, encore moins nombreux, les contrats souscrits de 2018. Il n'y a aucune saisonnalité. Il n'y a naturellement que très peu de sinistres concernant les contrats de 2019.

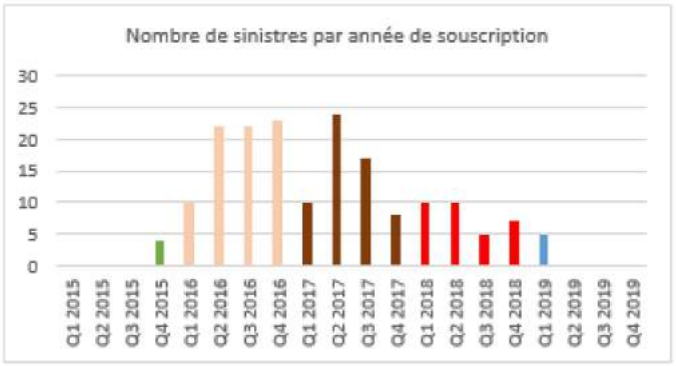


FIGURE 2.10 – Nombre de sinistres en fonction de la date de souscription

2015	4
2016	77
2017	59
2018	32
2019	5

TABLE 2.8 – Nombre de sinistres par année de souscription

Nous étudions ensuite les dates d'occurrence des sinistres.
Aucun sinistre n'a eu lieu en 2015. Ceux-ci ont débutés en 2016, très faiblement. De 2016 à 2018 le nombre de sinistres croît. Nous notons une augmentation assez importante du nombre de sinistres à partir du dernier trimestre de 2017. Le nombre de sinistre par trimestre double, de 10 à 20. Il y a un pic de sinistralité en début 2019. Il est essentiel de remarquer que nous avons probablement des sinistres IBNR (*incured but not reported*) sur les dernières années, notamment 2019. Ce sont des sinistres déjà survenus mais dont l'assureur n'a pas encore connaissance.

Comme nous l'avons évoqué précédemment, nous rappelons une des limites de notre base de données : son historique. Nous ne pouvons observer que cinq années d'historique sur le portefeuille global des partenaires, un délai relativement court par rapport à la durée moyenne des contrats de 9 ans. Ainsi nous ne prenons pas en compte une génération complète. Cela pourrait conduire à omettre certaines informations importantes qui pourraient affecter les taux de sinistralité. Aucune adaptation de la base à ce niveau ni vérification des hypothèses n'est possible, il faut donc faire très attention au niveau initial de prudence des hypothèses.

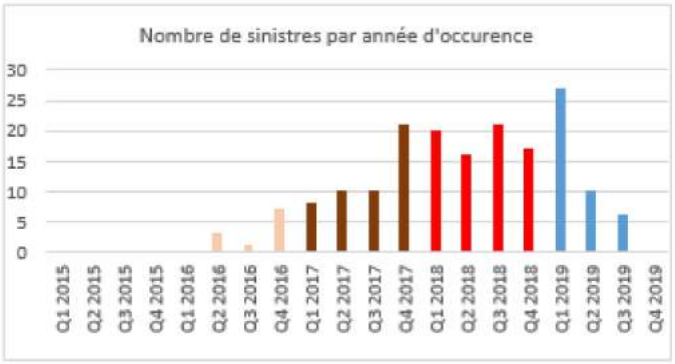


FIGURE 2.11 – Nombre de sinistres en fonction de la date d'occurence

2015	0
2016	11
2017	49
2018	74
2019	43

TABLE 2.9 – Nombre de sinistres par année d'occurence

La nature de la répartition de la catégorie socio-professionnelle

Nous observons parmi les employés ayant perdu leur emploi une très grande majorité d'employés appartenant à la branche Privé, soit 87% des sinistres. C'est un pourcentage relativement important compte tenu de la répartition des emprunteurs. Ainsi, nous pouvons d'ores et déjà prédire une très forte influence de la catégorie socio-professionnelle sur le taux d'incidence chômage. C'est un constat tout à fait cohérent : dans un contexte économique difficile, où les emplois précaires et les périodes de chômage se multiplient, la fonction publique représente une valeur moins incertaine. En effet, un fonctionnaire est titulaire d'un grade, ce qui signifie qu'il ne peut être rétrogradé ou licencié sans motif grave. De plus, en cas de restructuration d'une administration, il se voit proposer un nouvel emploi. Les risques de perte d'emploi sont donc considérablement amoindris.

Du fait de cette importante disparité, nous n'hésiterons pas, lors de nos diverses modélisations à départager notre base de données selon la catégorie socio professionnelle.

FIGURE 2.12 – Répartition des sinistres selon la catégorie Socio-professionelle

Répartition de l'âge des sinistrés

Nous observons alors, en fonction de l'âge à la souscription de contrat, l'occurrence des sinistres. L'âge moyen, pour les personnes ayant perdu leur emploi, est de 44 ans.
Les sinistrés sont très dispersés, d'âges compris entre 22 et 66 ans. Bien que la tendance globale soit normale, il y a beaucoup d'irrégularités avec notamment un pic important autour de 55 ans.

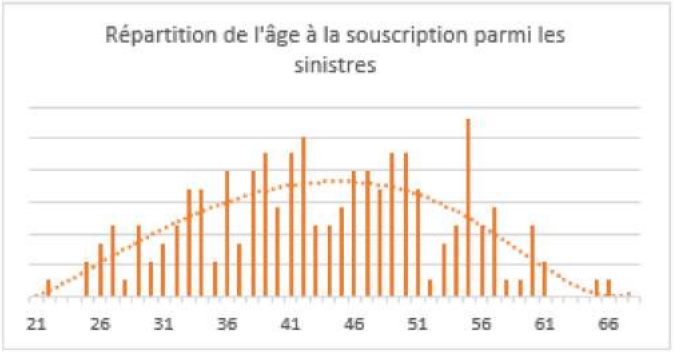


FIGURE 2.13 – Répartition de l'âge à la souscription parmi les sinistres

count	177
Mean	44
Std	9
min	22
25%	36
50%	43
75%	50
Max	66

TABLE 2.10 – Statistiques descrptives de l'âge à la souscritption parmi les sinistrés

Cela nous amène à nous interroger non plus sur la répartition de l'âge à la souscription, mais sur

la répartition de l'âge au moment des sinistres.
Au moment des sinistres, les assurés sont âgés de 24 à 68 ans. La répartition une fois encore est assez irrégulière, avec un pic important à l'âge de 50ans.

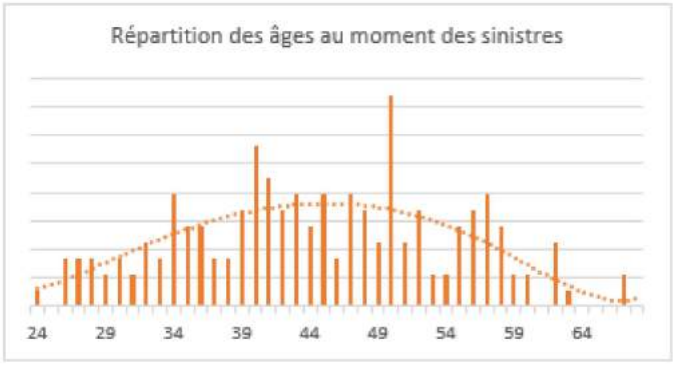


FIGURE 2.14 – Répartition des âges au moment des sinistres

Niveau des montants assurés

Alors que les montants assurés sont très étendus valent entre 1 900€ et 85 000€, nous remarquons que des sinistres ont eu lieu seulement pour des montants assurés valant entre 2900€ à 32 000€. Ainsi, contrairement à ce que nous pourrions penser, le risque d'insolvabilité de l'emprunteur suite à une perte d'emploi n'augmente pas du tout avec le montant de l'emprunt, au contraire, les sinistres concernent les emprunteurs de montants faibles. Le montant moyen initial étant de 12 900€.

count	177
Mean	12 880
Std	6 028
min	2 985
25%	8 023
50%	12 106
75%	16 491
Max	31 248

TABLE 2.11 – Statistiques descriptives des montants assurés parmi les sinistrés

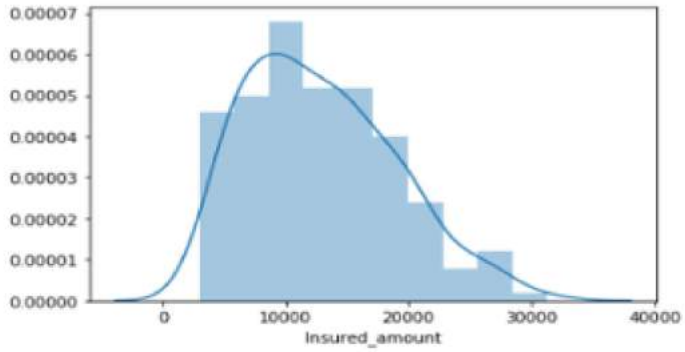


FIGURE 2.15 – Fonction de répartition des montants assurés parmi les sinistrés

Proportion homme/femme

Parmi les personnes ayant subi la perte de leur emploi, la proportion d'hommes est importante. Elle est de 77% tandis que la proportion d'hommes parmi les assurés est de seulement 62%. Il semblerait donc que les hommes de notre portefeuille soient plus enclins à la perte d'emploi que les femmes.

FIGURE 2.16 – Répartition du sexe parmi les sinistrés

Le partenaire financier octroyant le prêt

Nous rappelons que nous regroupons notre base en deux grandes catégories de partenaires : le Partenaire1 et les autres. La prépondérance du Partenaire1 (85% des emprunteurs) est due à son ancienneté supérieure aux autres. Ainsi ce-dernier est naturellement très important parmi les sinistrés (95%).

Bien qu'il s'agisse d'un nombre très élevé et qui semble relativement parlant, étant donné l'inégalité d'historique dont nous disposons, nous nous questionnerons par la suite sur le fait que ces valeurs sont, ou non, exploitables. Pour l'instant, nous n'emmétons aucune hypothèse quant au risque de l'une des catégorie par rapport à l'autre.

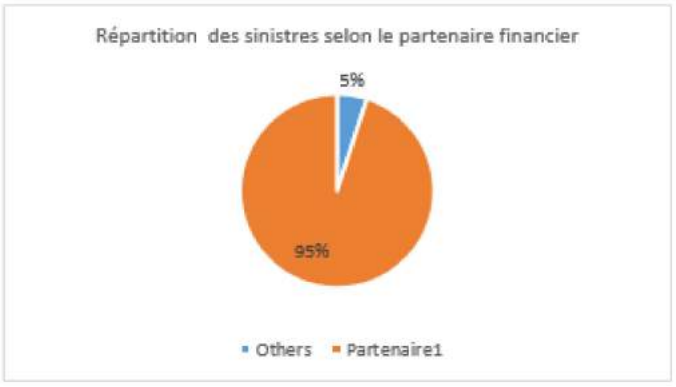


FIGURE 2.17 – Répartition des sinistres selon le partenaire financier

Répartition des produits

Parmi les enregistrements des sinistres, 19% concernent des prêts *Delega* et 81% concernent des prêts CQS. Nous rappelons que le *Delega* est un prêt complémentaire au CQS, et, en cas d'insolvabilité, le TFR est utilisé en premier lieu pour le remboursement du CQS.
Un calcul simple nous donne que donc 76% des sinistrés n'ont qu'un prêt CQS et 23% ont un prêt CQS et un prêt *Delega*. Au vu de ces résultat, il est pour le moment difficile d'emmettre une hypothèse quant à l'influence du produit.

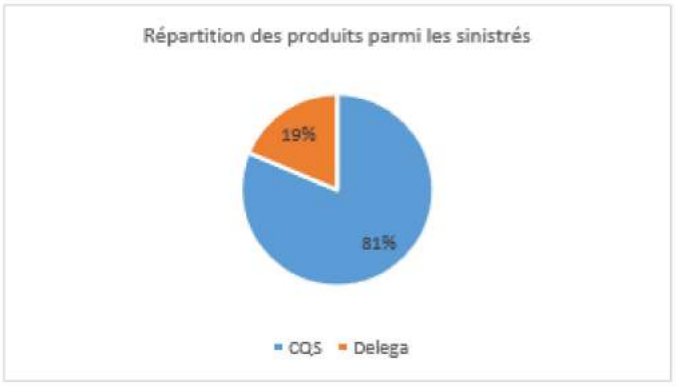


FIGURE 2.18 – Répartition des produits parmi les sinistrés

La durée du contrat

La durée moyenne des contrats est de 88 mois donc assez faible au vu de la durée moyenne des contrats sur l'ensemble du portefeuille.
Nous comparons, parmi les sinistrés, les durées initiales des contrats selon les catégories socio-professionnelles. La durée est nettement plus élevée pour le parapublic puis très proche pour les secteurs privés et publics.

count	177
Mean	88
Std	30
min	24
25%	60
50%	84
75%	120
Max	120

TABLE 2.12 – Statistiques descrptives de la durée des contrats parmi les sinistrés (en mois)

Privé	87
Para-public	108
Public	90

TABLE 2.13 – Durée moyenne des contrats parmi les sinistrés selon la catégorie socio-professionelle(en mois)

2.2.3 Analyse bivariée : corrélation entre les variables

Dans cette étape, nous cherchons à croiser nos variables explicatives et à mesurer leur corrélation. En effet, notre choix de modélisation implique l'absence de liaisons trop fortes entre les variables explicatives.

Liaison entre deux variables quantitatives

Deux coefficients sont utilisés afin de déterminer la relation entre deux variables quantitatives : le coefficient de Pearson et le coefficient de Spearman.

Le coefficient de Pearson : Il permet d'analyser les relations linéaires, mais n'est applicable que pour mesurer la relation entre deux variables dont la distribution et gaussienne. Le coefficient de corrélation linéaire de Pearson entre deux variables aléatoires réelles X et Y ayant chacune une variance finie est définie par :

$r = \frac{\text{Cov}(X,Y)}{\sigma_X \times \sigma_Y}$ où :
— $\text{Cov}(X, Y)$ désigne la covariance des variables X et Y
— σ_X et σ_Y désignent leurs écarts types
Si r est proche de 0, il n'y a pas de corrélation entre les deux variables. Si il est proche de -1, il existe une corrélation linéaire négative entre X et Y et si r est proche de 1, alors il existe une corrélation positive entre X et Y.

Le coefficient de Spearman : Il est une mesure de dépendance statistique non paramétrique entre deux variables.

La corrélation de Spearman est étudiée lorsque deux variables statistiques semblent corrélées sans que la relation entre les deux variables soit de type affine. Elle consiste à trouver un coefficient de corrélation, non pas entre les valeurs prises par les deux variables mais entre les rangs de ces valeurs. Elle estime à quel point la relation entre deux variables peut être décrite par une fonction monotone. Pour un échantillon de taille n, les variables de rangs $rg(X_i)$ et $rg(Y_i)$ sont calculées à partir des données X_i et Y_i.

La corrélation de Spearman est définie par :

$r_s = \frac{\text{cov(rg(X),rg(Y))}}{\sigma_{\text{rg(X)}} \times \sigma_{rg(Y)}}$ où :

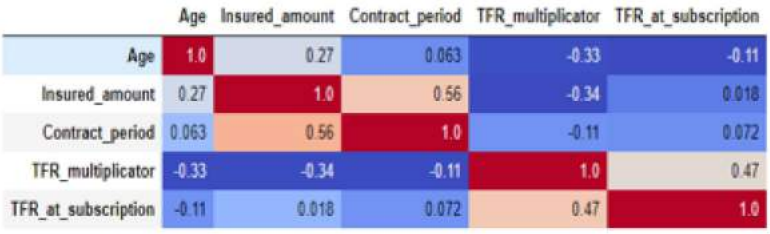
— $\text{cov}\,(\text{rg(X)}, \text{rg(Y)})$ désigne la covariance de variables de rang
— $\sigma_{\text{rg(X)}}$ et $\sigma_{\text{rg(Y)}}$ désignent les écarts types des variables de rang
On constate que cette définition correspond à la corrélation de Pearson des variables de rang.

S'il n'y a pas de données répétées, une corrélation de Spearman parfaite de +1 ou -1 est obtenue quand l'une des variables est une fonction monotone parfaite de l'autre.
Pour les deux coefficients, si le signe du coefficient est positif, cela signifie que les variables évoluent dans le même sens, dans le cas contraire, cela signifie qu'elles évoluent dans un sens opposé.
Notons néanmoins une différence entre causalité et liaison : ce n'est pas parce qu'il y a une liaison statistique entre deux variables qu'il y a forcément un lien de causalité.

Résultats obtenus : D'après les tableaux suivants, il n'y a aucune corrélation très forte entre les variables explicatives quantitatives.
La matrice de corrélation de Spearman notamment révèle une corrélation très élevée entre le coefficient multiplicateur du TFR et le montant du TFR lui-même. En effet, ces deux variables dépendent toutes deux fortement de l'employeur de l'emprunteur.
Les matrices révèlent aussi une corrélation non négligeable entre la durée du contrat et le montant assuré. C'est une constatation tout à fait logique.

	Age	Insured_amount	Contract_period	TFR_multiplicator	TFR_at_subscription
Age	1.0	0.25	0.021	-0.35	-0.33
Insured_amount	0.25	1.0	0.62	-0.38	-0.27
Contract_period	0.021	0.62	1.0	-0.12	-0.058
TFR_multiplicator	-0.35	-0.38	-0.12	1.0	0.77
TFR_at_subscription	-0.33	-0.27	-0.058	0.77	1.0

FIGURE 2.19 – Matrice de corrélation de Spearman



	Age	Insured_amount	Contract_period	TFR_multiplicator	TFR_at_subscription
Age	1.0	0.27	0.063	-0.33	-0.11
Insured_amount	0.27	1.0	0.56	-0.34	0.018
Contract_period	0.063	0.56	1.0	-0.11	0.072
TFR_multiplicator	-0.33	-0.34	-0.11	1.0	0.47
TFR_at_subscription	-0.11	0.018	0.072	0.47	1.0

FIGURE 2.20 – Matrice de corrélation de Pearson

Liaison entre deux variables qualitatives

Pour étudier les liaisons entre nos variables qualitatives, nous utilisons le V de Cramer. Afin de bien comprendre la statistique du V de Cramer, nous introduisons la statistique du χ^2 (test d'indépendance de deux caractères) car il en est directement dérivé.

Le test du χ^2 d'indépendance : Il a pour but d'évaluer si deux variables qualitatives X_1 et X_2 à respectivement p et k modalités sont liées, les deux variables étant observées sur un échantillon de taille N. Ce test se base sur la notion d'effectifs synthétisés au sein d'un tableau de contingence, qui comporte autant de lignes que de modalités de la variable X et autant de colonnes que de modalités de la variable X. A chaque croisement de la modalité i de X_1 avec la modalité j de X_2 est associé l'effectif observé o_{ij}. L'effectif observé de la modalité i de X_1 est noté t_i et l'effectif observé de la modalité j de X_2 est noté n_j.

Ce test nécessite le calcul des effectifs théoriques selon la formule : $e_{ij} = \frac{n_j t_i}{N}$.
Celà nous permet d'obtenir le tableau de contingence théorique suivant :

	Variable X_2				
Variable X_1	Modalité 1	Modalité 2	...	Modalité k	Total
Modalité 1	e_{11}	e_{12}	.	e_{1k}	t_1
Modalité 2	e_{21}	e_{22}	.	e_{2k}	t_2
.	.	.	.	.	
Modalité p	e_{p1}	e_{p2}	.	e_{pk}	t_k
Total	n_1	n_2	.	n_k	N

FIGURE 2.21 – Tableau de contingence théorique

L'unique condition d'application stipule que les effectifs théoriques doivent être supérieurs ou égaux à 5. Si ce n'est pas le cas, on procède à un regroupement de modalités.

Les hypothèses du test sont les suivantes :

— $H0$: Les variables X_1 et X_2 sont indépendantes

— $H1$: Il existe une liaison entre X_1 et X_2

Sous $H0$, la statistique de test associée au test du χ^2 d'indépendance est définie par :
$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{k} \frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(p-1)(k-1)ddl}$

Pour un risque de première espèce α, la région critique conduisant au rejet de $H0$ est définie par :

$W = \left[\chi^2_{(1-\alpha);(p-1)(k-1)ddl}\,; +\infty\right]$ où :

— $\chi^2_{(1-\alpha);(p-1)(k-1)ddl}$ correspond au quantile d'ordre $(1-\alpha)$ de la loi du χ^2 à $(p-1)(k-1)$ dégrés de liberté

Si la valeur de la statistique de test est inférieure au quantile, alors on conserve l'hypothèse nulle, les variables X_1 et X_2 sont indépendantes. Si la valeur de la statistique de test est supérieure, on rejette l'hypothèse nulle, il existe une liaison significative entre X_1 et X_2.

Le V de Cramer : Contrairement au χ^2, le V de Cramer reste stable si l'on augmente la taille de l'échantillon dans les mêmes proportions intermodalités. Il est basé sur le χ^2 maximal que le tableau de contingence pourrait théoriquement produire : ce dernier aurait alors une seule case non nulle par ligne ou par colonne (selon que le tableau a plus de lignes ou plus de colonnes). Ce χ^2 maximum théorique est égal à l'effectif multiplié par le plus petit côté du tableau (nombre de lignes ou de colonnes) moins 1. Le V de Cramer est la racine carrée du χ^2 divisée par le $\chi^2 max$:

$$V = \sqrt{\frac{\chi^2}{\chi^2\,\max}} = \sqrt{\frac{\chi^2}{n\times[\min(l,c)-1]}}$$

Plus V est proche de zéro, plus il y a indépendance entre les deux variables étudiées. Il vaut 1 en cas de complète dépendance entre les variables.

	Sex	Name_of_partner	Type_of_product	Socio_professional_category
Sex	1.0	0.0	0.0	0.03
Name_of_partner	0.0	1.0	0.01	0.01
Type_of_product	0.0	0.01	1.0	0.02
Socio_professional_category	0.03	0.01	0.02	1.0

FIGURE 2.22 – Matrice de Cramer

Les liaisons entre les variables explicatives sont négligeables.

Finalement, nos diverses analyses sur les variables qualitatives et quantitatives, nous mènent à la conclusion qu'il existe effectivement des corrélations entre les variables explicatives, mais celles-ci peuvent être considérées comme faibles.

2.3 Première étude et analyse du taux d'incidence perte d'emploi

Nous avons effectué, dans la partie précédente, une première analyse statistique de nos variables explicatives. Désormais nous allons calculer le taux d'entrée annuel au chômage de notre portefeuille.

2.3.1 Quelques définitions

Nous rappelons que la proportion de sinistres dans une population à un instant donné, sans distinction entre les anciens sinistres et les nouveaux, est appelée prévalence.

Elle est définie par la formule : $P(t) = \frac{m(t)}{n(t)+m(t)}$ avec :
- $m(t)$ est le nombre d'assurés ayant subis un sinistre
- $n(t)$ le nombre d'assurés n'ayant pas subis de sinistre

C'est une mesure instantanée. La prévalence de la sinistralité de notre base est de 1.75%, ce qui est relativement faible.

L'incidence correspond au nombre de nouveaux cas de sinistres ou d'assurés ayant subi un sinistre pendant une période déterminée $[t; t + dt]$. Ainsi, nous obtenons la formule suivante :

$$\text{Incidence} = \frac{\text{Nombre de nouveaux sinistres pendant la période d'étude}}{\text{Nombre de personnes * temps à risque pendant la période d'étude}}$$

D'où le taux d'incidence sur une période se calcule de la façon suivante :

$$\text{Taux} = \frac{\text{Nombre de décès observés}}{\text{Nombre de contrats * Temps exposés}}$$

Dans notre cas, toutes les dates sont connues.

Résultats obtenus : Par application de la formule ci-dessus, nous calculons les taux d'incidence chômage de notre portefeuille. A cette étape, nous le calculons selon deux granulations : le partenaire financier et la catégorie socio-professionnelle des assurés en portefeuille.

Nous choisissons de classer les sinistres par date de survenance, et non par date de déclaration. Prenons l'exemple d'un sinistre survenu en novembre 2016 mais déclaré en janvier 2017, ce sinistre sera rattaché à l'année 2016.

Nous rappelons que le taux d'acceptation des sinistres est de 100%, il n'est donc pas nécessaire de s'en préoccuper dans nos calculs.

Le Partenaire1 est le plus ancien partenaire sur le produit CQS.
Les autres partenariats n'ont débuté qu'en 2018 voire 2019 pour certains. De ce fait, les résultats pour les autres partenaires ne débutent qu'à cette date et son probablement bien moins pertinents.

Les taux obtenus sont relativement faibles, quelle que soit la catégorie socio-professionnelle et le partenaire financier, avec un maximum de 2,05%. Comme nous l'avons présenté dans la Partie 1, les CQS sont un produit en plein essor et très appréciés notamment du fait d'une forte rentabilité due à des taux de sinistralités très faibles.

	PUB	PARA	PRIV	Total
2015	0,00%	0,00%	0,00%	0,00%
2016	0,11%	1,60%	2,38%	0,83%
2017	0,25%	0,69%	3,37%	1,43%
2018	0,23%	0,49%	2,89%	1,33%
2019	0,00%	0,00%	0,00%	0,00%
Total	0,19%	0,40%	2,00%	0,93%

TABLE 2.14 – Taux d'incidence chômage sur l'ensemble du portefeuille

	PUB	PARA	PRIV	Total
2015	-	-	-	-
2016	-	-	-	-
2017	-	-	-	-
2018	0,00%	-	0,00%	0,00%
2019	0,26%	0,00%	1,36%	0,86%
Total	0,26%	0,00%	1,33%	0,84%

TABLE 2.15 – Taux d'incidence chômage des contrats avec les partenaires *Others*

	PUB	PARA	PRIV	Total
2015	0,00%	0,00%	0,00%	0,00%
2016	0,11%	1,60%	2,38%	0,83%
2017	0,25%	0,69%	3,37%	1,43%
2018	0,23%	0,49%	2,90%	1,33%
2019	0,15%	0,00%	0,90%	0,46%
Total	0.19%	0,43%	2,05%	0,46%

TABLE 2.16 – Taux d'incidence chômage des contrats avec le Partenaire1

Nous remarquons que le taux est le moins élevé pour les employés du secteur publique.Les employés du secteur parapublique viennent ensuite avec un taux deux fois plus élevé. Concernant le secteur privé, le taux est environ 10 fois plus élevé que pour les fonctionnaires. C'est un constat tout à fait cohérent avec ce que nous avions notifié lors de nos statistiques descriptives. Le secteur public est caractérisé par une certaine sécurité de l'emploi. Le licenciement tout comme la démission se font sous des conditions très réglementées, contrairement au secteur privé ou ces procédures sont plus facilement accessibles.

Il est important de remarquer que les taux observés sur l'année 2019 sont relativement plus faibles que les années précédentes. Ceci nous laisse penser que nous ne disposons pas de l'ensemble de l'information et nous allons donc nous intéresser dans la partie suivante aux sinistres qui sont survenus, mais dont l'assureur n'en a pas encore connaissance.

2.3.2 Estimation des sinistres inconnus

Nous étudions la répartition des sinistres en fonction de leur délai de déclaration, c'est-à-dire la différence entre la date à laquelle les sinistres ont effectivement lieu et la date à laquelle l'assureur en prend connaissance. L'objectif est de déterminer si nous observons, dans notre base, tous les sinistres, ou bien si les sinistres IBNR (*Incured but not reported* - provision pour sinistres inconnus) sont nombreux auquel cas, nous allons redresser les taux d'entrée au chômage calculés dans la partie précédente.
Le fait que nous n'ayons que très peu d'historique rend l'impact des IBNR d'autant plus important.

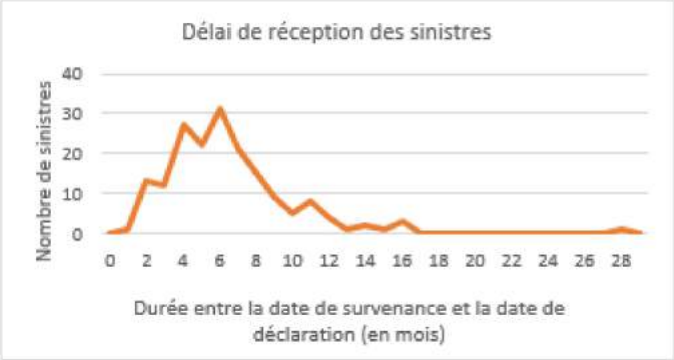


FIGURE 2.23 – Délai de réception des sinistres

La majorité des sinistres sont déclarés entre 0 et 17 mois. Au-delà de 6 mois, le nombre de sinistres déclarés diminue progressivement. Le délai de déclaration moyen est de 6 mois. Les délais de déclaration sont donc non négligeables.

La méthode Chain Ladder Avec la méthode de Clain ladder nous allons estimer les IBNR sur notre période d'observation, à savoir survenus avant le 01 décembre 2019 mais non déclarés au 01 décembre 2019.

Nous avons créé des triangles de déclaration, en reportant, pour chaque année de survenance et chaque année de déclaration le nombre de sinistres déclarés. Puis, nous appliquons la méthode de

Chain Ladder.

La méthode de Chain Ladder est une méthode déterministe. Elle est souvent utilisée car elle est comprise et mise en place très facilement. Nous l'appliquons à des triangles de sinistres cumulés.

Nous considérons une périodicité annuelle. Nous utilisons les notations suivantes :
— i l'année de survenance des sinistres
— j l'année de déclaration des sinistres
— $(X_{i,j})$ les sinistres incrémentaux
— $C_{i,j}$ les sinistres cumulés tel que : $C_{i,j} = \sum_{k=1}^{j} X_{i,k}$
D'où le tableau ci-dessous, dont nous cherchons à estimer la partie blanche, correspondant aux sinistres survenus avant l'année n.

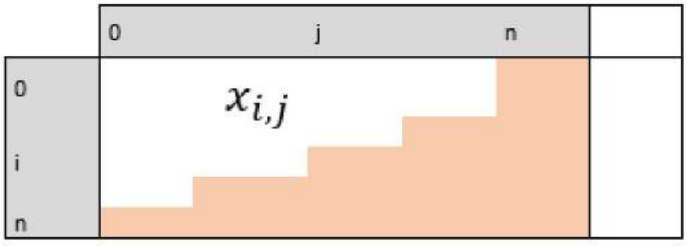


FIGURE 2.24 – Tableau utilisé pour la méthode Chain Ladder

La méthode Chain Ladder repose sur l'hypothèse suivante : Pour j allant de 1 à n, les facteurs de développement $f_{i,j}$ sont indépendants de l'année de survenance i.

Les coefficients de passage, d'une année à l'autre, communs pour les années de survenance sont estimés par :

$$f_j = \frac{\sum_{i=0}^{n-j+1} C_{i,j+1}}{\sum_{i=0}^{n-j+1} C_{i,j}}, j = 0, \ldots, n$$

Résultats obtenus Grâce à ces facteurs, nous pouvons estimer le nombre de sinistres survenus avant l'année n.

Sur l'ensemble du portefeuille, nous obtenons le triangle ci-dessous :

	2016	**2017**	**2018**	**2019**
2016	3	11	11	11
2017	19	47	49	**49**
2018	23	74	**77**	**77**
2019	43	**126**	**130**	**130**

TABLE 2.17 – Application de la méthode Chain Ladder a notre portefeuille

Nous appliquons la méthode et calculons des IBNRs. Nous obtenons les résultats suivants sur l'ensemble du portefeuille.

2016	0
2017	0
2018	3
2019	87

TABLE 2.18 – Montant des IBNRs calculés avec Chain Ladder

Seules les années 2018 et 2019 comptent des IBNRs.
Alors, nous appliquons de même cette méthode selon la granularité précédente, c'est-à-dire selon la catégorie socio-professionnelle et le partenaire financier afin d'obtenir une nouvelle estimation des taux d'incidence.

Il nous est impossible d'effectuer la méthode Chain Ladder sur les autres partenaires, car nous ne disposons pas d'un historique de données suffisant.

	PUB	**PARA**	**PRIV**	**Total**
2015	0,00%	0,00%	0,00%	0,00%
2016	0,11%	1,60%	2,38%	0,83%
2017	0,25%	0,69%	3,37%	1,43%
2018	0,68%	0,98%	3,02%	1,65%
2019	0,49%	0,90%	2,65%	0,45%
Total	0,47%	0,94%	2,86%	1,46%

TABLE 2.19 – Taux d'incidence chômage sur l'ensemble du portefeuille (en considérant les IBNRs)

	PUB	**PARA**	**PRIV**	**Total**
2015	0,00%	0,00%	0,00%	0,00%
2016	0,11%	1,60%	2,38%	0,83%
2017	0,25%	0,69%	3,37%	1,43%
2018	0,68%	0,98%	3,03%	1,66%
2019	0,46%	1,03%	2,48%	1,33%
Total	0.45%	0,99%	2,81%	1,41%

TABLE 2.20 – Taux d'incidence chômage des contrats avec le Partenaire1 (en considérant les IBNRs)

Les valeurs calculées avec les IBNRs diffèrent ainsi des valeurs calculées à la partie précédente. En effet les taux d'incidence sur 2019 semblent bien plus alignés à ceux des années précédentes.

Nous disposons d'une base de données présentant les contrats souscrits et une autre présentant les sinistres survenus. Ces deux bases ont été mises en commun, analysées et nettoyées pour être utilisées par la suite.
Nous calculons, en appliquant la définition mathématique, le taux d'incidence perte d'emploi selon la catégorie socio-professionelle et les années. Nous adaptons alors les résultats en incluant les IBNRs, calculés avec une méthode Chain Ladder. Ceux-ci ninfluent de façon non négligeables sur les taux d'incidence, notamment sur l'année 2019.
Nous faisons une première observation : les taux sont relativement plus importants pour le secteur

privé que pour les deux autres secteurs. Ainsi, la catégorie socio-professionelle semble être une variable très influente.

Chapitre 3

Modélisation avec les méthodes d'analyse de survie

Cette partie a pour but de présenter les premiers outils théoriques d'ores et déjà à notre disposition pour l'élaboration des taux d'incidence de la perte d'emploi. Nous présentons plusieurs méthodes puis précisons celle que nous retenons pour le calcul des taux d'entrée au chômage annuel segmentés selon les différentes variables explicatives. Enfin,, nous définissons les variables les plus discriminantes ainsi que les taux d'entrée finalement retenus.

L'analyse de survie a pour objectif d'analyser et de modéliser les données et notamment la survenance d'un événement d'intérêt. En actuariat, l'événement d'intérêt peut être le décès, mais aussi l'apparition d'une maladie ou la survenue d'un sinistre. Lors de notre étude l'événement d'intérêt sera en effet la survenue de la perte d'emploi. Nous tentons d'en comprendre la cause et d'établir les facteurs de risque.
En analyse de survie, les données ont toujours les trois particularités suivantes :

— La variable à expliquer est le temps d'attente jusqu'à la survenue de l'événement : la durée de survie, qui est toujours positive ou nulle.
— Les observations sont censurées. Nous verrons dans le paragraphe suivant ce que cela signifie.
— Nous disposons de variables explicatives dont nous cherchons l'influence sur la durée de survie.

Nous utilisons, de manière analogique, les méthodes d'analyse de survie pour le calcul des taux de chômage. La durée de vie correspond à la durée de vie des contrats, le décès correspond à l'incidence de l'évènement cible à savoir la perte d'emploi. La survie correspond au maintien de l'emploi.

3.1 L'analyse de survie

3.1.1 Quelques définitions

Nous utiliserons dans la suite le vocabulaire relatif à l'analyse de survie.

— La naissance correspond à l'entrée en portefeuille.
— Le décès correspond à la survenance de l'événement d'intérêt (par exemple la survenue d'un sinistre, ici le sinistre perte d'emploi).
— La survie correspond à l'absence de survenance de l'événement d'intérêt.
— La durée de survie désigne le temps écoulé jusqu'à la survenance d'un événement . L'analyse

des données de survie permet de l'étudier.
Ainsi, nous cherchons à estimer les distributions des temps de survie (fonction de survie), nous comparons celles des différents groupes, afin d'analyser l'influence des variables explicatives.

Certains termes sont à connaître :
- La date d'origine correspond à l'origine de la durée étudiée
- La date de point correspond à la date au-delà de laquelle on cesse l'étude

On suppose que la durée de survie X est une variable positive ou nulle et absolument continue.

Elle peut être définie par, pour t fixé :
Notons $X \geq 0$, alors :
- Sa fonction de survie : variable aléatoire modélisant la durée de survie, supposée absolument continue. Elle est la probabilité de survivre jusqu'à t : $S(t) = P(X > t)\,;\, t \geq 0$
  Elle diminue avec le temps et sa valeur initiale est 1.
- Sa fonction de répartition : c'est la probabilité de mourir avant t : $F(t) = P(X \leq t) = 1 - S(t)$
- Sa densité de probabilité : c'est la probabilité de mourir dans un petit intervalle de temps après t. C'est la fonction $f(t) \geq 0$ telle que, pour tout $t \geq 0$, $F(t) = \int_0^t f(u)du$
  si F est dérivable en t, $f(t) = \lim_{h\to 0} \frac{\mathbb{P}(t\leq X<t+h)}{h} = F'(t) = -S'(t)$
- Le risque instantané λ (ou taux de hasard) : Il s'agit d'une autre fonction, avec la fonction de survie, caractéristique de la distribution de X. C'est la probabilité de mourir dans un petit intervalle de temps après t sachant que l'on a survécu jusqu'à t :
  $\lambda(t) = \lim_{h\to 0} \frac{\mathbb{P}(t\leq X<t+h|X\geq t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))'$
  Le numérateur est la probabilité conditionnelle que le décès se produise dans l'intervalle $[t; t+h[$ sachant qu'il ne s'est pas produit avant et le dénominateur est la taille de l'intervalle.
- Le taux de hasard cumulé Λ, c'est l'intégrale du risque instantané : $\Lambda(t) = \int_0^t \lambda(u)du = -\ln(S(t))$
  On a : $S(t) = \exp(-\Lambda(t)) = exp\left(-\int_0^t \lambda(u)du\right)$
  d'où : $f(t) = \lambda(t)\exp\left(-\int_0^t \lambda(u)du\right)$

Les quantités associées à la distribution de survie sont :
- Le temps moyen de survie : $\mathbb{E}(X) = \int_0^\infty S(t)dt$
- La durée de survie : $\mathbb{V}(X) = 2\int_0^\infty tS(t)dt - (\mathbb{E}(X))^2$
- Les quantiles : le quantile $q(p)$ est le temps où une proportion p de la population a disparu. la médiane de la durée de survie est le temps t_m tel que $S(t_m) = 0,5$.
  Pour les estimateurs qui sont des fonctions en escaliers (tels que Kaplan Meier) il est possible qu'un intervalle de temps vérifie cette équation. Il faut donc être très vigilant pour l'interprétation.

  La fonction quantile de la durée de survie est définie par :
  $q(p) = \inf(t : F(t) \geq p)$
  $\quad = \inf(t : S(t) \leq 1 - p)$ avec $0 < p < 1$

  Lorsque la fonction de répartition F est strictement croissante et continue, alors nous avons :
  $q(p) = F^{-1}(p)$
  $\quad = s^{-1}(1 - p)$ avec $0 < p < 1$

3.1.2 Phénomènes de censure et de troncature

La censure

La deuxième caractéristique des données de survie est la présence du phénomène de censure. Lors de l'étude d'un problème d'analyse de survie, il est possible que l'événement d'intérêt ne soit pas observé dans certains cas : ce scénario se produit en raison de la fenêtre temporelle d'observation limitée ou d'éléments manquants. Ce concept est connu sous le nom de censure. Elle est très souvent rencontrée. Soit un individu i, on note X_i son temps de survie, C_i son temps de censure et T_i la durée réellement observée.

- La censure à droite : elle a lieu si l'individu n'a pas subi l'événement à la date de la dernière observation. Ainsi, nous savons simplement que la durée de vie est supérieure à une valeur donnée.
  - Censure de type I : Soit C une valeur fixée, au lieu d'observer les variables $X_1.., X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i < C$; sinon on sait uniquement que $X_i > C$. On pose $T_i = X_i \wedge C = min(X_i; C)$
  - Censure de type II : elle est présente lorsque l'on décide d'observer les durées de survie de n individus jusqu'à ce que k d'entre eux subissent l'évènement et d'arrêter l'étude à ce moment.
  - Censure de type III (ou aléatoire de type I) . Soient $C_1, \dots C_n$ des variables *i.i.d.* Nous observons les variables $T_i = min(Xi; Ci)$. Ainsi, l'information disponible se résume par :
    - La durée réellement observée T_i
    - Un indicateur $\delta_i = \mathbf{1}_{\{X_i \leq C_i\}}$ qui vaut 1 si l'événement est observé et 0 sinon (cas des données incomplètes, censurées).

    Ce type de censure est le plus courant
- La censure à gauche : elle a lieu si l'individu a déjà subi l'événement avant le début de l'observation. Ainsi, nous savons simplement que la durée de vie est inférieure à une valeur donnée. On peut associer, à chaque individu, un couple de variables aléatoires (T, δ) :
  $T = X \vee C = \max(X, C)$
  $\delta = \mathbf{1}_{\{X \geq C\}}$
  Nous nous intéresserons, dans notre étude, uniquement à la censure à droite.
- La censure par intervalle : elle a lieu si on ne connaît que l'intervalle de temps pendant lequel a eu lieu l'événement.

Dans les trois cas, le temps d'occurrence réel de l'événement est inconnu.

La date de censure correspond au minimum entre la date de fin d'observation et la date de sortie du chômage. La date de sortie du risque chômage correspond au minimum entre la date de fin de crédit et l'âge limite de la couverture.

Le graphique suivant illustre la censure à droite.

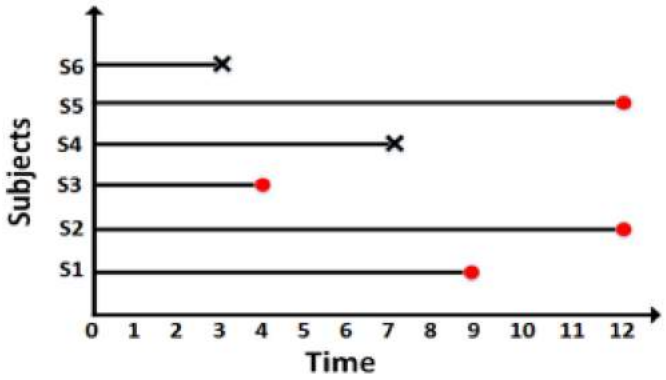


FIGURE 3.1 – Représentation graphique de la censure à droite

Pour un problème de survie, le temps qui s'écoule avant l'événement est connu avec précision lorsque celui-ci a lieu. Dans les autres cas, nous ne connaissons que le temps censuré. Ainsi, pour un individu donné i nous ne pouvons observer que le temps de survie ou le temps de censure, mais pas les deux. Nous considérons que la censure est une variable aléatoire indépendante de l'événement. Cette hypothèse est indispensable pour l'utilisation des modèles classique d'analyse de survie. Dans notre étude, la censure est causée par la fin de l'étude, cette hypothèse est donc vraie.

Présentation du problème Pour une instance i donnée, représentée par un triplet (X_i, y_i, δ_i), où Xi est le vecteur des caractéristique ; δi est l'indicateur d'événement binaire, et y_i indique le temps observé.
On a donc : $y = T_i$ si $\delta_i = 1$ et $y = C_i$ si $\delta_i = 0$.

Ainsi, L'objectif de l'analyse de survie est d'estimer la durée avant l'événement d'intérêt T_j pour un nouvel individu j avec les variables caractéristiques de l'individu X_j.

La troncature

Contrairement à la censure, la troncature est liée à l'échantillonnage. Une variable X est tronquée par un sous ensemble A de $\mathbf{R}^+$, si, au lieu de X, nous observons seulement X si $X \in A$. Ainsi, les événements de l'échantillon appartiennent tous à A et suivent une loi T conditionnée par cette appartenance. Ainsi, s'il y a troncature, nous n'étudions qu'une partie de la population. Soit Z une variable aléatoire indépendante de X, il y a une troncature à gauche si X est observable seulement si $X > Z$. De même, il y a une troncature à droite si X est observable seulement si $X < Z$. Il y a une troncature par intervalle si X est tronquée à droite et à gauche.

La date de troncature correspond au maximum entre la date de début d'observation et la date d'entrée au risque chômage. Le cas de troncature est donc le début de l'observation.

3.2 Les méthodes de modélisation

Les méthodes statistiques usuelles peuvent être subdivisées en trois catégories :

— Les modèles non paramétriques
— Les modèles semi-paramétriques
— Les modèles paramétriques

Les méthodes non paramétriques sont plus efficaces lorsqu'il n'y a pas de distribution sous-jacente pour l'occurrence de l'événement ou lorsque l'hypothèse de risque proportionnel ne tient pas. Parmi les méthodes non paramétriques la fonction de survie peut être obtenu avec la méthode de l'estimateur de Kaplan Meier ou encore la méthode de l'estimateur de Nelson Aalen.

Parmi les modèles semi-paramétriques, le modèle de Cox est le plus couramment utilisé pour l'analyse des durées de survie. Contrairement aux méthodes précédentes, il est basé sur l'hypothèse des risques proportionnels et utilise des estimations partielles des paramètres. Ce modèle est appelé semi-paramétrique car la distribution des résultats reste finalement inconnue.

Finalement, les méthodes paramétriques sont plus précises et efficaces lorsque la durée avant l'événement d'intérêt suit une distribution particulière spécifiée par certains paramètres. La méthode de régression linéaire est l'une des principales méthodes paramétriques utilisée en analyse de survie.

3.2.1 Les modèles non paramétriques

Comme indiqué précédemment, au vu de nos données, nous nous plaçons dans le cas d'une censure à droite aléatoire de type 1.

Estimateur de Kaplan Meier de la fonction de survie

L'estimateur de Kaplan Meier est l'un des principaux estimateurs non paramétriques de la fonction de survie. Cet estimateur doit son nom à Edward L. Kaplan et Paul Meier. L'avantage important de cet estimateur est que c'est une méthode permettant de prendre en compte les données censurées, notamment censurées à droite, ce qui, comme nous l'avons vu précédemment, intervient lorsque nous ne disposons plus des données de certains individus avant que l'événement d'intérêt ne soit survenu.

Celui-ci vient du concept suivant : nous disposons de deux temps t et t' tels que $t' < t$ alors la probabilité de survivre au-delà de t peut s'écrire :

$$\begin{aligned} S(t) &= P[T > t] = P[T > t', T > t] \\ &= P[T > t | T > t'] \times P[T > t'] \\ &= P(T > t | T > t') \times S(t') \end{aligned}$$

Nous listons (T_i, i de 1 à n) l'ensemble des temps censurés de notre échantillon rangés par ordre croissant. On choisit comme instant de conditionnement les instants où se produisent une sortie ou censure. Cela nous ramène alors à estimer des probabilités de la forme :

$$pi = \mathbf{P}(T > T_i | T > T_{i-1}) \text{ avec } P(T > T_i) = \prod_{i=1}^{n} p_i$$

pi est la probabilité de survivre sur l'intervalle $]T_{i-1}; T_i]$ sachant qu'on était vivant à l'instant T_{i-1}. Nous choisirons la convention $T_0 = 0$

Un estimateur naturel de $q_i = 1 - p_i$ est $\hat{q}_i = \frac{di}{ri}$ avec :

— d_i le nombre personnes décédées en T_i
— r_i le nombre d'individus à risque de subir l'évènement juste avant le temps T_i

Comme les temps d'évènements sont supposés distincts, $d_i = 0$ en cas de censure en T_i *i.e.* si $\delta i = 0$ et $d_i = 1$ en cas de décès en T_i *i.e.* $\delta i = 1$.

Alors, nous obtenons l'estimateur de la fonction de survie Kaplan Meier :

$$\begin{aligned}\hat{S}(t) &= \prod_{i=1,T_i<t}^{n} 1-\frac{\delta_i}{r_i} \\ &= \prod_{i=1,T_i<t}^{n} 1-\frac{\delta i}{n-(i-1)} \\ &= \prod_{i=1,T_i<t}^{n} [\frac{n-i}{n-i+1}]^{\delta_i}\end{aligned}$$

L'estimateur de Kaplan Meier est aussi appelé produit limite car il est défini comme la limite d'un produit.
La fonction obtenue est une fonction décroissante, en escalier et continue à droite.
Nous pouvons montrer qu'il s'agit d'un estimateur du maximum de vraisemblance.

Sous certaines hypothèses, lorsque le nombre d'individus à risque est élevé, il est possible de démontrer que cet estimateur est uniformément consistant et asymptotiquement normal.

Dans le cas où il y a des ex-aequos :
— Si ce sont des évènements de nature différente, on considère que les observations non censurées ont lieu avant les censurées.
— Si il y a plusieurs décès au même temps T_i, alors $d_i > 1$ et on a l'estimation suivante : $\hat{S}(t) = \prod_{i=1,T_i<t}^{n} 1-\frac{di}{ri}$

Il existe des estimateurs de la variance de $\hat{S}(t)$, dont l'estimateur de Greenwood (Klein, 1991) :

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i,T_{(i)}\leq t} \frac{d_i}{r_i(r_i-d_i)}$$

Cette estimation est obtenue en partant de l'approximation suivante :

$$\widehat{\mathrm{Var}}(\log(\hat{S}(t))) = \sum_{i,T_{(i)}\leq t} \frac{d_i}{r_i(r_i-d_i)}$$

On utilise la Delta méthode : $\mathrm{Var}(f(Z)) \approx [f'(E(Z))]^2 \mathrm{Var}(Z)$ pour montrer que :

$$\widehat{\mathrm{Var}}(\log(\hat{S}(t))) \approx \frac{1}{\hat{S}(t)^2} \mathrm{Var}(\hat{S}(t))$$

Déduction des taux bruts

A partir de l'estimateur de la fonction de survie, nous pouvons aisément déduire les taux bruts de mortalité via la relation :

$$\begin{aligned}\hat{q}_x &= 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)} \\ &= 1 - \frac{\prod_{i|a_i<x+1} \frac{r_i-d_i}{r_i}}{\prod_{i|a_i<x} \frac{r_i-d_i}{r_i}} \\ &= 1 - \prod_{i=a_1}^{a_m} \frac{r_i - d_i}{r_i}\end{aligned}$$

Dans chaque intervalle de temps, l'estimation de la fonction de survie est une proportion. Sous certaines conditions, nous pouvons faire une approximation par la loi normale et ainsi obtenir un intervalle de confiance :

$$IC(\alpha) = [\hat{S}(t) \pm z_{\frac{\alpha}{2}} \sqrt{\hat{\text{Var}}(\hat{S}(t))}]$$

Nous ne pouvons pas utiliser cet intervalle quand $\hat{S}(t)$ est proche de 0 ou de 1. Car, l'intervalle étant symétrique autour de $\hat{S}(t)$, les bornes peuvent dépasser les valeurs 0 ou 1. C'est pourquoi nous préférons utiliser l'intervalle de confiance de Rothman (1978) :

$$IC(\alpha) = \frac{K}{K+\left(z_{\frac{\alpha}{2}}\right)^2}[\hat{S}(t) + \frac{\left(z_{\frac{\alpha}{2}}\right)^2}{2K} \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{S}(t)) + \frac{\left(z_{\frac{\alpha}{2}}\right)^2}{4K^2}}] \text{ avec } K = \frac{\hat{S}(t)(1-\hat{S}(t))}{\hat{\text{Var}}(\hat{S}(t))}$$

3.2.2 Les modèles semi-paramétriques

Le modèle de Cox est la méthode d'analyse de survie la plus couramment utilisée parmi les modèles semi-paramétriques. Nous nous restreindrons donc à la présentation de celle-ci.

Le modèle de Cox

Le modèle de Cox, introduit par le statisticien David Cox, est aussi connu sous le nom d'analyse de régression des risques proportionnels. Il permet d'étudier la relation entre la survie d'un individu et plusieurs variables explicatives. Ce modèle permet de modéliser la survenance de l'événement d'intérêt en prenant en compte les particularités du portefeuille mais permet aussi de considérer les phénomènes de censure, qui, comme nous l'avons expliqué précédemment, sont très fréquemment rencontrés dans les problématiques d'analyse de survie.
Contrairement aux méthodes paramétriques la connaissance de la distribution sous-jacente du temps jusqu'à l'événement d'intérêt n'est pas nécessaire mais les variables sont supposées avoir une influence exponentielle sur le résultat.

Le modèle de Cox permet d'exprimer la fonction de risque instantané λ, c'est-à-dire le risque instantané de décès à l'instant t sachant que l'individu est vivant juste avant t, en fonction du temps t et des variables explicatives $X_1, \ldots, X_n$.

Nous avons alors $\lambda(t, X_1, \cdots, X_n) = \lambda_0(t) \exp(\Sigma_{i=1}^n \beta_i X_i)$ avec : $\beta = (\beta_1, \ldots, \beta_n)$ le vecteur des paramètres et $X = (X_1, \ldots, X_n)$ le vecteur des n variables explicatives (pouvant d'ailleurs dépendre du temps).

Dans cette formule, $\lambda_0(t)$ est appelé la fonction de hasard de base. Il correspond au risque instantané de décès lorsque toutes les variables sont nulles. Ainsi, il ne dépend ni des variables explicatives, ni du vecteur de paramètres β mais seulement du temps.
La deuxième partie $\exp(\Sigma_{i=1}^{n}\beta_i X_i)$ ne dépend quant à elle que des variables. Ce modèle est appelé semi-paramétrique car on ne cherche pas à estimer $\lambda_0(t)$, mais le rapport des risques instantanés de décès pour des individus exposés différemment.

Il découle de cette formule une hypothèse primordiale pour le modèle de Cox : l'hypothèse de proportionnalité des risques et l'hypothèse de log linéarité.

Risques proportionnels

Soit j_1 et j_2 deux individus qui ne qui ne diffèrent que par la valeur de la k-ième variable, alors pour tout t :

$$\frac{\lambda(t,j_2)}{\lambda(t,j_1)} = \frac{\lambda_0(t)\exp\left(\beta_1 X'_1+...+\beta_{k-1}X'_{k-1}+\beta_k\times 1+\beta_{k+1}X'_{k+1}+...+\beta_n X'_n\right)}{\lambda_0(t)\exp\left(\beta_1 X'_1+...+\beta_{k-1}X'_{k-1}+\beta_k\times 0+\beta_{k+1}X'_{k+1}+...+\beta_n X'_n\right)} = \exp(\beta_k)$$

Cette hypothèse signifie que le rapport, en logarithme, entre les courbes de hasard de deux groupes d'individus est non seulement proportionnel à la différence entre les valeurs de la variable, mais surtout indépendant du temps.

Log-linéarité

$\log(h(t|X_{i1},...,X_{ip})) = \log(h_0(t)) + B_0 Z_i$
Le logarithme du risque instantané est une fonction linéaire des Z_{ij}.

Détermination des paramètres du modèle

Pour estimer les B_k(k=1 à n) nous utiliserons la méthode du maximum de vraissemblance à l'aide de l'algorithme de Newton Raphson. Pour un échantillon de données indépendantes, nous posons les notations suivantes :

- t_i avec i= 1 à m les temps de décès
- d_i le nombre de décès au temps t_i
- D_i l'ensemble des décès au temps t_i
- r_i le nombre d'individus a risque de mourir en t_{i-}
- R_i l'ensemble des individus à risque de mourir en t_{i-}

En t_i, la probabilité qu'un sujet k décède sachant qu'il est vivant en t_{i-} est : $\lambda_0(t_i)\exp(X_k^t\beta)$

Au temps t_i, la vraisemblance de l'ensemble des individus appartenant à D_i vaut : $\frac{\exp\left(\beta^t\Sigma_{k\in D_i}X_k\right)}{\left[\Sigma_{l\in R_i}\exp(\beta^t X_l)\right]^{d_i}}$

Finalement, la vraisemenblance (partielle) du modèle peut s'écrire :
$L(\beta) = \prod_{i=1}^{m}\frac{\exp\left(\beta^t\Sigma_{k\in D_i}X_k\right)}{\left[\Sigma_{l\in R_i}\exp(\beta^t X_l)\right]^{d_i}}$

Pour un échantillon de taille fixe, Anderson et Grill (1982) ont prouvé la convergence de l'estimateur

du maximum de vraisemblance sous certaines conditions de régularité, et qu'il est asymptotiquement normal. Ainsi nous avons la convergence en loi : $\sqrt{n}\left(\hat{B}^{MV} - B\right) \overset{n\to\infty}{\longrightarrow} N\left(0, \sum^{-1}\right)$

Interprétation du modèle en terme de rapport de risques

Dans un modèle de Cox comme nous l'avons présenté, nous disposons à la fois de variables discrètes et de variables continues. L'interprétation du paramètre dépend du type de variable auquel celui-ci est associé.

Pour une variable X_k discrète qui ne prend que deux valeur 0 ou 1 (X_k est l'indicatrice d'un évènement), comme par exemple le sexe que nous choisirons à titre indicatif.
X_{ik}= {0 si l'individu i est un homme et 1 si individu i est une femme }

Nous avons alors :

$$\frac{h_0(t)\exp(\beta_1 Z_{i1}+\ldots+\beta_k\times 1+\ldots+\beta_p Z_{ip})}{h_0(t)\exp(\beta_1 Z_{i1}+\ldots+\beta_k\times 0+\ldots+\beta_p Z_{ip})} = \exp(\beta_k)$$

Alors, $\exp(B_k)$ est le rapport de risque entre une femme et un homme, toutes choses étant égales par ailleurs (c'est-à-dire après avoir pris en compte la valeur des autres variables).

— Si $B_k > 0$ c'est à dire $\exp(B_k) > 1$ alors le risque de décès est plus élevé chez les femmes que chez les hommes
— Si $B_k < 0$ c'est à dire $\exp(B_k) < 1$ alors le risque de décès est plus élevé chez les hommes que chez les femmes
— Si $B_k = 0$ c'est à dire $\exp(B_k) = 1$ alors le risque de décès est le même chez les hommes et les femmes.

Pour une variable X_k continue, le coefficient associé B_k verifie :

$$Bk = \frac{\partial \log(ho(\frac{t}{X}))}{\partial Xk}$$

Donc B_k est l'élasticité du taux de hasard par rapport à la k-ième variable continue X_k supposée constante par rapport au temps. Donc B_k est l'effet de la variable Z_k sur le risque instantané toutes choses étant égales par ailleurs. Il peut être vu comme le rapport de risque quand la variable associée X_k augmente d'une unité.

$$\frac{h_0(t)\exp(\beta_1 Z_{i1}+\ldots+\beta_k Z_{ik+1}+\ldots+\beta_p Z_{ip})}{h_0(t)\exp(\beta_1 Z_{i1}+\ldots+\beta_k Z_{ik}+\ldots+\beta_p Z_{ip})} = \exp(\beta_k)$$

— Si $B_k > 0$ c'est-à-dire $\exp(B_k) > 1$ alors le risque de décès augmente quand Zk augmente (resp. diminie qd Zk diminue)
— Si $B_k < 0$ c'est-à-dire $\exp(B_k) < 1$ alors le risque de décès diminue quand Zk augmente (resp. augmente quand Zk diminue)
— Si $B_k = 0$ c'est-à-dire $\exp(B_k) = 1$ alors la variable Xk n'a pas d'impact significatif sur le risque décès

Validation du modèle

Le modèle de Cox postule une hypothèse majeure, que les risques sont proportionnels entre individus. Nous devons vérifier cette hypothèse, afin de nous assurer de la fiabilité des résultats. Plusieurs approches sont possibles :

— méthode graphique : comparaison des courbes de survie
— analyse des résidus de Schoenfeld
— test de la non-interaction avec le temps

Si l'hypothèse des risques proportionnels n'est pas vérifiée pour une variable du modèle nous pouvons nous en affranchir, en stratifiant le risque de base par rapport à cette variable.

Comparaison graphique Nous considérons la transformation suivante des courbes de survie : $\log(-\log(S(t))$. Cette transformation, dite LML (*Log minus log*) a la propriété suivante : Si l'hypothèse des risques proportionnels est valide, alors nous avons :

$$S(t,x) = S_0(t)exp(X'\beta) \Leftrightarrow \log(-\log S(t,x))=\log(-\log(S_0(t))) + X'\beta$$

Pour une variable X_k qui prend les valeurs X_1 ou X_2, la différence entre les courbes LML vaut $(X_2' - X_1')\beta$. Cette quantité est indépendante du temps.
Ainsi, les courbes de survie après transformation LML sont parallèles pour différentes modalités de x. Nous traçons alors les courbes LML correspondant aux différentes modalités de la variable, les autres covariables restant constantes, et nous les comparons. S'il est possible de superposer les différentes courbes par simple translation, alors l'hypothèse de proportionnalité est vérifiée.

Pour une variable continue, nous devons tout d'abord répartir ses valeurs en un nombre fini de catégories. Il n'existe pas de règles précise définissant si les courbes sont similaires ou non, une part de subjectivité intervient.

Analyse des résidus Usuellement, dans un modèle de régression linéaire par exemple, les résidus correspondent à la différence entre les valeurs observées de la variable étudiée et les valeurs prédites par le modèle.
Avec les résidus de Schoenfeld, pour chaque date de mort nous calculons la différence entre les caractéristiques de l'individu décédé (en cas d'ex-aequo, nous calculons un résidu pour chaque individu et chaque temps de décès et sommons les résidus) et une moyenne pondérée des caractéristiques des individus à risque de décéder au temps t_i.

Nous obtenons alors : $R_{ij} = X_{ij} - \bar{X}_{ij}(t_i)$
avec :

— R_{ij} : résidu au temps t_i
— X_{ij} : valeur de la covariable j pour l'individu décédé au temps t_i
— $\bar{X}_{ij}(t_i)$: moyenne pondérée de la covariable j chez les individus à risque au temps t_i

Nous utilisons les résidus standardisés, c'est-à-dire les résidus divisés par leur variance.

Les résidus de Schoenfeld peuvent être analysés sur la base de graphiques, afin de détecter un éventuel non-respect de l'hypothèse de proportionnalité. L'idée consiste à représenter les résidus en fonction d'une transformation du temps. Nous pouvons ajouter, sur le même graphique une courbe représentant l'évolution moyenne des résidus en fonction du temps. Cette courbe donne la tendance générale.

Si l'hypothèse est bien vérifiée, l'ensemble des résidus doivent être distribués de la même manière au cours du temps. Ainsi, toute différence par rapport à une droite horizontale représente une déviation par rapport à l'hypothèse de proportionnalité.

Coefficients de régression dépendant du temps Une autre méthode de test consiste alors à introduire des coefficients de régression évoluant en fonction du temps dans le modèle de Cox et à tester leur significativité. S'ils sont significatifs, alors l'hypothèse de proportionnalité des risques est remise en question.

3.2.3 Les modèles paramétriques

Les modèles paramétriques supposent que le temps de survie ou le logarithme du temps de survie suivent une distribution particulière. Ces modèles sont simples, efficaces et performants. Les distributions couramment utilisées dans les modèles de régression paramétrique censurés sont : normale, exponentielle, Weibull, logistique, log-logistique et log-normale. Si les temps de survie de tous les individus suivent ces distributions, le modèle est appelé modèle de régression linéaire. Si le logarithme des temps de survie de tous les individus suit ces distributions, le modèle peut aussi très bien être analysé. Il convient de noter que si aucune distribution théorique appropriée n'est connue, les méthodes non paramétriques sont plus efficaces.

3.3 Les méthodes de lissage

Après avoir estimé les taux bruts de survenance du chômage dans notre portefeuille, qui présentent certaines inégalités dues à l'imperfection des conditions de l'expérience, il est nécessaire de procéder à un ajustement, ou lissage de nos valeurs brute, afin de présenter de manière plus fidèle la loi que l'on souhaite estimer. Le choix de la méthode la plus adéquate prend en compte deux critères essentiels :

— La précision, aussi appelée fidélité : les taux réajustés doivent être le plus proches des taux initiaux
— La régularité : la suite des taux lissés doit être aussi régulière que possible

Nous distinguons deux types de modèles de lissages. Les modèles paramétriques sont basés sur l'utilisation d'une loi sous-jacente usuelle à déterminer. Les méthodes non paramétriques, quant à elles, ne reposent sur aucune famille de loi.

3.3.1 Les lissages paramétriques

Le lissage paramétrique consiste à déterminer, parmi une famille de fonction statistique, la plus adaptée à nos données. Nous présenterons uniquement les lissages par la méthode de Spline et de Gompertz- Makeham. Dans le cas où la variable à expliquer est une fonction linéaire des variables explicatives, nous utilisons la régression adaptée aux modèles linéaires généralisé, et, quand ce n'est pas le cas, nous utilisons la régression non linéaire. Celle-ci est une méthode itérative qui repose sur un choix de valeurs initiales pour les paramètres, choix qui sera très important pour la suite, afin de nous assurer d'une convergence de la procédure de régression.

La méthode des Spline

L'idée du lissage par Spline est de partager la plage de la fonction à ajuster en plusieurs sous-intervalles, puis, sur chacun d'entre eux, ajuster une fonction simple en faisant attention aux points de jonction. Plus le découpage est adapté, plus les fonctions utilisées sur chaque sous-intervalles sont simples.

Une fonction Spline de degré d (avec $d > 0$) à l noeuds $k_1,\ldots,k_l$ est une fonction de classe C^{d-1} dont les restrictions aux $[k_i\,;k_{i+1}]$ (avec i= 0 à l) sont des polynômes de degré d. Usuellement, nous choisissons $k_0 = -\infty$ et $k_{l+1} = +\infty$.

Le lissage d'une fonction Spline aux données brutes obtenues se fait avec la méthode des moindres carrés pondérés. Celle-ci consiste en la minimisation de la somme pondérée des écarts quadratiques entre les estimations brutes et ajustées.
Bien qu'il n'y aiT pas de règle afin de connaître le nombre de noeud optimal, nous notons que plus le nombre de noeuds est élevé, plus les valeurs lissées sont proches des estimations initiales. Dans l'idée de minimisation de la somme pondérée des écarts quadratiques, il peut par moment être plus pertinent de déplacer un noeud que d'en ajouter un.

Exemple de Spline cubique à deux arcs Nous supposons une partition de notre plage d'ancienetés en deux parties d'où :
$q_x = p_0(x)$ pour $x_0 \leq x \leq x_1$ et $q_x = p_1(x)$ pour $x_1 \leq x \leq x_2$ avec p_0 et p_1 des polynômes de degré 3.
D'où les contraintes aux points de jonction :
$p_0\left(x_1\right) = p_1\left(x_1\right)\,;\, \frac{d}{dx}p_0\left(x_1\right) = \frac{d}{dx}p_1\left(x_1\right)\,;\, \frac{d^2}{dx^2}p_0\left(x_1\right) = \frac{d^2}{dx^2}p_1\left(x_1\right)$

Nous déduisons donc :
$p_0(x) = c_1 + c_2x + c_3x^2 + c_4x^3$ et $p_1(x) = p_0(x) + c_5\left(x - x_1\right)^3$

En retirant les 3 contraintes de régularité aux 8 coefficients des polynômes, cela nous ramène à un problème à 5 inconnues. Alors, nous procédons à la méthodedes moindres carrés sur la base des poids w_x, nous souhaitons minimiser : $M = \sum_{x=x_0}^{x_2} w_x\left(q_x - \hat{q}_x\right)^2$

En cas de valeurs manquantes, le Spline a la capacité de les interpoler.
On note alors $\bar{x_1}$ la plus grande valeur de x inférieure ou égale à x_1 pour laquelle on dipose d'une valeur $\hat{q_x}$, on décompose la somme intervenant dans le critère M en deux sommes puis on écrit les équations normales en annulant les dérivées par rapport aux paramètres : $\frac{\partial M}{\partial c_i} = 0$

Après calcul, ces équations peuvent se mettre sous la forme : $X'wXc = X'w\hat{q}$

La matrice X de taille $(m, 5)$ pour m valeurs de $\hat{q_x}$ disponibles sur $[x_0, x_2]$ étant définie par :

$$\mathrm{X} = \begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3 & 0 \\ .. & .. & .. & .. & .. \\ 1 & \bar{x_1} & \bar{x_1}^2 & \bar{x_1}^3 & 0 \\ 1 & \bar{x_1}^1 & (\bar{x_1}^1)^2 & (\bar{x_1}^1)^3 & (\bar{x_1}^1 - x_1)^3 \\ .. & .. & .. & .. & .. \\ 1 & x_2 & x_2^2 & x_2^3 & (x_2 - x_1)^3 \end{pmatrix}$$

avec $\bar{x_1}^1$ la valeur de l'indice postérieure à $\bar{x_1}$ pour laquelle $\hat{q_x}$ est connue.

3.3.2 Les lissages non - paramétriques

Nous présenterons parmi les méthodes de lissage non paramétrique, la méthode des moyennes mobiles pondérées et celle de Whittaker-Henderson.

Les moyennes mobiles

La méthode des moyennes mobiles, l'une des premières méthodes de lissage développée, a l'avantage d'être très simple à mettre en oeuvre. Elle présente néanmoins des faiblesses aux bornes des intervalles de lissage, liées notamment à la sensibilité de la moyenne aux valeurs extrêmes. Ainsi, cette méthode n'est pas préférée aujourd'hui.

Nous nous restreindrons aux moyennes mobiles symétriques d'expression :
$q_x = \sum_{i=-r}^{+r} a_i \hat{q}_{x+i}$ avec $a_{-i} = a_i$

Pour diminuer l'erreur de prédiction, nous pouvons fixer des contraintes : si la série des taux bruts q_x semble assimilable à un polynôme et ainsi régulière, alors, nous imposons la contrainte que l'application de la moyenne mobile ne modifie pas les valeurs brutes.
C'est-à-dire, avec l'expression précédente : $\sum_{i=-r}^{+r} a_i = 1$ et $\sum_{i=-r}^{+r} i^2 a_i = 0$

Cette méthode, est souvent mal adaptée et conduit à des irrégularités. Nous nous concentrons donc davantage sur la méthode ci-après, qui, au contraire, est très utilisée, car elle donne de très bons résultats.

La méthode de Whittaker-Henderson

Le principe de cette méthode est d'ajuster les taux en minimisant la combinaison linéaire de deux critères :

— Le critère de fidélité : $F = \sum_{i=1}^{p} w_i \left(q_i - \hat{q}_i\right)^2$
— Le critère de régularité : $S = \sum_{i=1}^{p-z} \left(\Delta^z q_i\right)^2$ avec z un paramètre du modèle.

Nous souhaitons donc minimiser $M = F + h \cdot S$ qui satisfait aux conditions $\frac{\partial M}{\partial q_i} = 0$ avec $1 \leq i \leq p$

Nous posons alors :
— $q = (q_i)_{1 \leq i \leq p}$
— $\hat{q} = (\hat{q}_i)_{1 \leq i \leq p}$
— $w = diag(w_i)_{1 \leq i \leq p}$
Pour la fidélité nous avons donc : $F = (q - \hat{q})' w (q - \hat{q})$ et pour la régularité, avec $\Delta^z q = (\Delta^z q_i)_{1 \leq i \leq p-z}$ alors $S = (\Delta^z q)' \Delta^z q$.

Nous introduisons alors la matrice K de taille $(p - z, p)$ dont les termes sont les coefficients binomiaux d'ordre z de signe alterné et qui commence positivement pour z pair :

$$\Delta^z q(i) = \sum_{j=0}^{2} \begin{pmatrix} z \\ j \end{pmatrix} (-1)^{z-j} q(j+i)$$

Par exemple pour $z = 2$ et $p = 5$ on a :

$$\mathrm{K}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix}$$

Nous pouvons aisément étendre ce modèle en dimension.
Les estimations sont $\hat{q} = (\hat{q}_{ij})_{1 \leq i \leq p, 1 \leq j \leq q}$ et le critère de fidélité devient :
$F = \sum_{i=1}^{p} \sum_{j=1}^{q} w_{ij} \left(q_{ij} - \hat{q}_{ij}\right)^2$.

Pour le critère de régularité, on calcule d'abord, via l'opérateur Δ_v^z la régularité verticale : $S_v = \sum_{i=1}^{p-z} \sum_{j=1}^{q} (\Delta_v^z q_{ij})^2$

Nous calculons ensuite la régularité horizontale S_h. Il faudra alors minimiser : $M = F + \alpha \cdot S_v + \beta \cdot S_h$

Pour ce, il sera nécessaire de se ramener au cas unidimensionnel. Ainsi, nous définissons un vecteur $p \cdot q$ tel que : $u_{q(i-1)+j} = \hat{q}_{ij}$. Ainsi pour les q premiers éléments du vecteur nous prenons la première ligne de la matrice $\hat{q}$ puis la deuxième ligne etc..

Nous obtenons aussi la matrice des poids en copiant sur la diagonale les lignes de la matrice (w_{ij}), d'où : $w^*_{q(i-1)+j, q(i-1)+j} = w_{ij}$.
De la même façon, nous obtenons K_z^v et K_y^h.

Nous déduisons alors les valeurs ajustées par la formule :
$q^* = (w^* + \alpha K_z^{v\prime} K_z^v + \beta K_y^{h\prime} K_y^h)^{-1} w^* u$

3.3.3 Validation du lissage

Afin de déterminer le modèle de lissage le plus adapté il existe de nombreuses métriques. Nous allons définir celles que nous utiliserons.

Le coefficient de détermination R^2

Il permet de mesurer la qualité du lissage. Il est défini par :

$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$ où :
— n est le nombre de mesures
— y_i la valeur de la i-ème mesure
— $\hat{y}_i$ la valeur prédite correspondante
— $\bar{y}$ la moyenne des mesures
Ce coefficient est le rapport de la variance expliquée sur la variance totale.

L'erreur absolue moyenne en pourcentage (Mean Absolute Percentage Error - MAPE)

Cette statistique correspond à la moyenne des écarts, en valeur absolues des valeurs brutes initialement calculées par rapport aux valeurs lissées. Il s'agit d'un pourcentage et donc, par conséquent, un indicateur pratique de comparaison. Néanmoins, elle présente l'inconvénient de ne s'appliquer qu'aux

valeurs strictement positives.

M $= \frac{1}{n}\sum_{t=1}^{n} \mid \frac{A_t - F_t}{A_t} \mid$ avec :

— A_t les valeur brutes
— F_t les valeurs lissées
Le modèle est d'autant plus adapté que ce pourcentage est faible. Plus le R^2 est élevé, plus nous privilégierons le modèle.

Le test d'adéquation du χ^2

Ce test permet de s'assurer qu'il n'y a pas de déformation non aléatoire du lissage. Nous disposons de n variables aléatoires X_i discrètes, à valeurs dans $\{a_1, \ldots, a_k\}$, indépendantes et de même loi caractérisée par $p = (p_1, \ldots, p_k)$ où $p_i = P(X = a_i)$.

On souhaite tester :
{ H_0=l'échantillon suit p_0 } contre { H_1 l'échantillon ne suit pas p_0 }
où $p_0 = (p_1^0, \ldots, p_k^0)$ est connue.

Nous utilisons à cet effet le test du χ^2 avec comme statistique : $\xi_n = n\sum_{i=1}^{k} \frac{\left(\hat{p}_i - p_i^0\right)^2}{p_i^0}$ $\hat{p}_i$ est la fréquence empirique de la valeur p_i.
La loi asymptotique de ξ_n sous H_0 est connue, il s'agit d'une loi du χ^2 à $k-1$ degrés de liberté.

Nous privilégierons le modèle avec la statistique la plus faible.

Le test de Kolmogorov-Smirnov

De même que le test précédent, ce test permet de mesurer la qualité du lissage.
Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d de même loi que X admettant F comme fonction de répartition.
On souhaite tester :
{ $H_0 : F = F_0$ } contre { $H_1 : F \neq F_0$ } où F_0 est connue.

On utilise à cet effet le test de Kolmogorov-Smirnov avec comme statistique :
$D_n = \sup_{x\in\mathbb{R}} |F_{emp}(x) - F_0(x)|$

où F_{emp} est la fonction de répartition empirique de l'échantillon telle que :
$F_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i} \leq x$

Cette statistique D_n est basée sur la distance entre fonction de répartition. Nous privilégierons le modèle avec la statistique la plus faible.

3.4 Construction et cadre d'applications

Dans cette section, nous allons construire les taux d'incidence Perte d'emploi de notre portefeuille à l'aide des méthodes présentées précédemment. Pour estimer la fonction de survie, nous utilisons

d'abord l'estimateur de Kaplan Meier, puis nous appliquons nos données au modèle de Cox.

Pour cette estimation, nous avons besoin de la durée écoulée de chaque contrat, s'ils ont été observés ou pas pendant toute la durée du contrat, pour l'enregistrement ou non d'un sinistre perte d'emploi.

Nous choisirons de développer les résultats pour les segmentations selon la catégorie socio-professionnelle, selon la répartition homme/femme et selon le produit.

3.4.1 L'estimateur de Kaplan Meier

Nous fixons la date de fin d'observation, donc la valeur maximum pour la date de censure au 30 novembre 2020.

Les courbes sont relativement irrégulières par nature : nous constatons beaucoup de fluctuations. La principale raison est le nombre de sinistres très peu élevé de notre portefeuille. Dans un premier temps, nous calculons les courbes de survie selon les différents critères. Nous choisissons donc de présenter les taux bruts calculés ainsi que le lissage le plus adapté.

Après avoir comparé, pour chacun des graphiques, pour chacune des méthodes de lissages divers critères de validations nous sélectionnons le lissage le plus adapté. Nous comparons les coefficients de détermination, l'erreur absolue moyenne et effectuons deux tests statistiques : le test d'adéquation du Chi2 et le test de Kolmogorov.

Résultats obtenus Pour chacune des courbes, la méthode de lissage la plus adaptée et cohérente avec nos résultats sera finalement la méthode Whittaker Henderson.

La catégorie socio-professionnelle

Dans un premier temps, nous choisissons de calculer les fonctions de survie selon la catégorie Socio-professionnelle.

La courbe de survie pour les employés du secteur privé est en bien dessous de celle des employés du secteur parapublique, elle-même en dessous mais proche de celle des employés du secteur public.

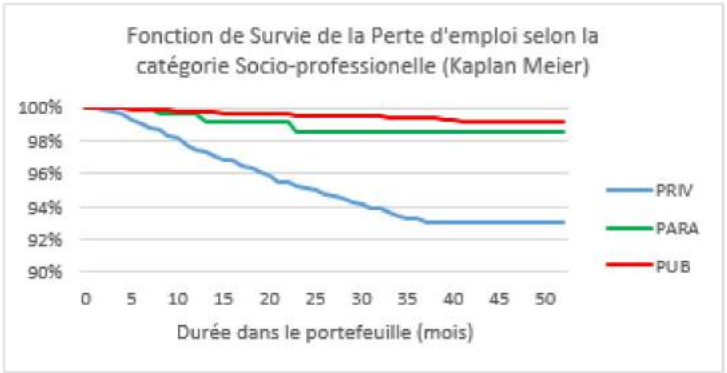


FIGURE 3.2 – Fonction de Survie de la Perte d'emploi selon la catégorie Socio-professionelle

Nous déduisons alors les taux d'incidence en fonction de la durée passée dans le portefeuille.

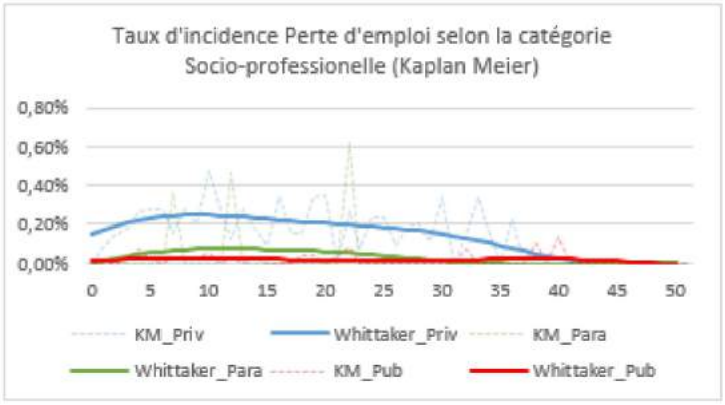


FIGURE 3.3 – Taux d'incidence perte d'emploi selon la catégorie Socio-professionelle

Ces résultats confirment nos premières conclusions concernant l'influence de la catégorie socio-professionnelle. Sur l'ensemble de la période, le taux d'incidence est significativement plus important pour le secteur privé. Ceci s'explique par les licenciements et démissions qui sont bien plus courants dans le secteur privé.

Le secteur public accorde à ses fonctionnaires une sécurité de l'emploi diminuant considérablement les risques de chômage.

Ce résultat est primordial, nous devons le prendre en compte lors de la détermination de la tarification. En effet, celle-ci diffère surtout selon le secteur d'activité de l'emprunteur.

La répartition homme - femme

Nous présentons les courbes de survie selon le sexe de l'emprunteur.

FIGURE 3.4 – Fonction de Survie de la Perte d'emploi selon le Sexe

D'après la courbe de survie, la population masculine est une population plus risquée que la

population féminine. Sa courbe de survie associée est située en-dessous.

Sur l'ensemble de la période étudiée, le taux d'incidence du chômage est plus élevé pour la population masculine que féminine. Néanmoins, nous notons que la tendance tend à s'inverser à partir du $34^{ème}$ mois, à partir duquel les taux d'incidence sont légèrement plus élevés pour les femmes que les hommes.

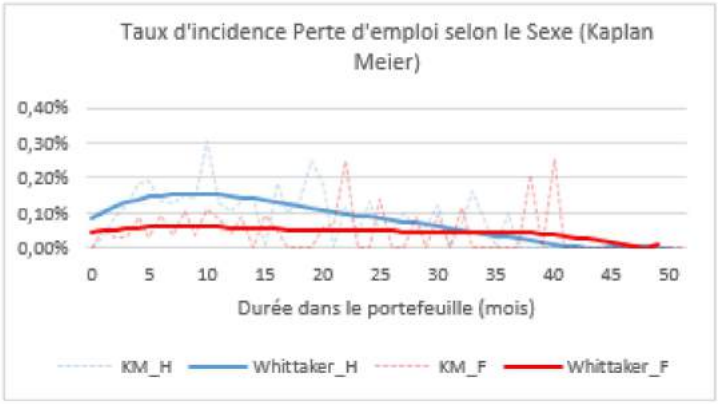


FIGURE 3.5 – Taux d'incidence Perte d'emploi selon le Sexe

Le type de produit

Nous calculons ensuite les fonctions de survie en fonction du type de produit. Nous nous demandons si les emprunteurs ayant à la fois un CQS et un *Delega* sont plus sujet à la perte d'emploi que les emprunteurs ayant uniquement un CQS. Nous analysons les enregistrements individuellement afin de comparer les fonctions de survie et taux d'incidence des deux produits.

La courbe de survie pour les *Delega* est globalement en-dessous de celle des CQS. Les deux courbes sont néanmoins très proches l'une de l'autre.

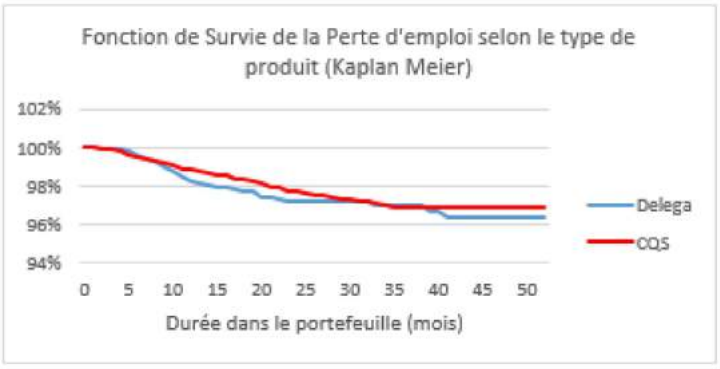


FIGURE 3.6 – Fonction de Survie de la Perte d'emploi selon le Type de Produit

Globalement les emprunteurs de prêts Delega semblent avoir un taux d'incidence Perte d'emploi plus élevé que les CQS uniquement, mais la tendance s'inverse entre le $12^{ème}$ et le $30^{ème}$ mois. Nous calculons le taux d'incidence moyen qui est de 0,06% pour les CQS, alors qu'il est de 0,08% pour les

Delega.

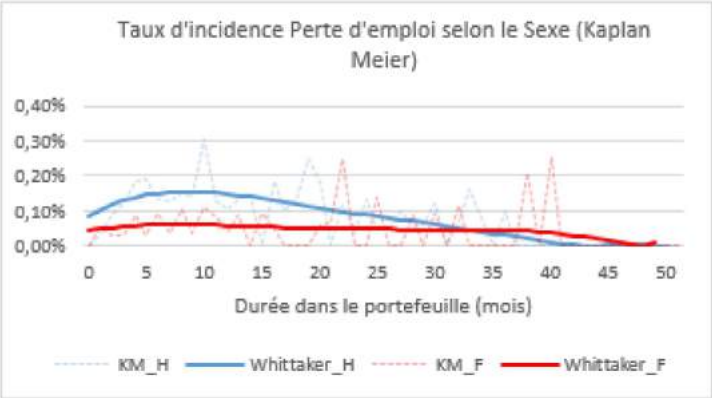


FIGURE 3.7 – Taux d'incidence Perte d'emploi selon le Sexe

3.4.2 La Méthode de Cox

Après avoir étudié l'impact des diverses variables sur la fonction de survie au risque chômage avec la méthode de Kaplan Meier, nous effectuons les mêmes analyses avec la méthode de Cox.

Une fois les hypothèses du modèle de Cox validées, à l'aide des modules Python, nous en estimons les paramètres et obtenons les résultats suivants.

model	lifelines.CoxPHFitter
duration col	'Duration'
event col	'Claim'
baseline estimation	breslow
number of observations	10087
number of events observed	177
partial log-likelihood	-1429.31
time fit was run	2020-05-07 11:28:10 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
Age	-0.02	0.98	0.01	-0.04	-0.00	0.96	1.00	-2.15	0.03	4.89
Insured_amount	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-2.15	0.03	4.89
Contract_period	-0.01	0.99	0.00	-0.01	0.00	0.99	1.00	-1.79	0.07	3.77
Sex_M	0.52	1.68	0.19	0.15	0.88	1.17	2.40	2.82	<0.005	7.66
Name_of_partner_OTHERS	-0.31	0.73	0.37	-1.04	0.42	0.35	1.52	-0.83	0.40	1.31
Type_of_product_DELEGA	0.60	1.83	0.20	0.22	0.99	1.24	2.69	3.06	<0.005	8.83
Socio_professional_category_PRIV	1.37	3.95	0.59	0.22	2.53	1.25	12.53	2.34	0.02	5.66
Socio_professional_category_PUB	-0.59	0.55	0.62	-1.80	0.63	0.16	1.88	-0.94	0.35	1.53

Concordance	0.60
Log-likelihood ratio test	218.66 on 8 df
-log2(p) of ll-ratio test	139.96

FIGURE 3.8 – Paramètres du modèle de Cox

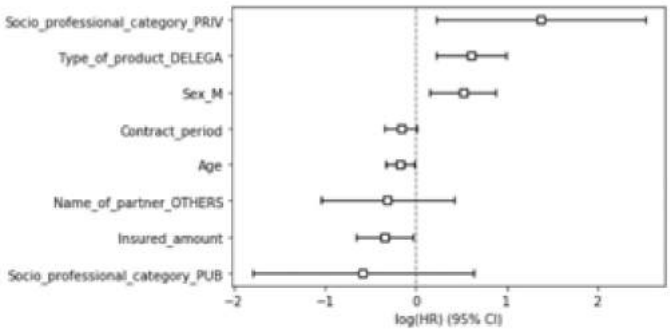


FIGURE 3.9 – Importance des covariables du modèle de Cox

Les statistiques sommaires ci-dessus indiquent l'importance des covariables dont nous disposons dans la prévision du risque Perte d'emploi. Ainsi, d'après ce modèle, les deux principales variables sont la catégorie socio-professionnelle et le nom du partenaire. Le type de produit, la répartition homme/femme et le montant assuré se positionnent ensuite. Enfin, l'âge et la durée du contrat semblent de faible importance.

Nous noterons que les résultats ci-dessus illustrent une importance du nom du partenaire dans la modélisation du risque. Néanmoins, comme nous disposons d'un faible historique sur les partenaires autres que le Partenaire1, ces résultats sont clairement biaisés. Nous ne tiendrons donc pas compte de l'influence de covariable pour la suite.

Comme dans la partie précédente, nous traçons les courbes de survie en ne faisant varier qu'une seule covariable tout en maintenant les autres égales. Compte tenu du modèle, ceci est utile pour comprendre l'impact d'une covariable, compte tenu du modèle.

La catégorie socio-professionnelle

FIGURE 3.10 – Fonction de Survie de la Perte d'emploi selon la catégorie Socio-Professionelle

Comme prévu, la fonction de survie est significativement inférieure pour le secteur privé. Puis très proche pour les secteurs parapublic et public, bien que légèrement inférieure pour le secteur parapublic. Les tendances des courbes sont les mêmes que celles obtenues avec la méthode de Kaplan Meier.

La répartition homme-femme

Nous comparons les fonctions de survie des hommes et des femmes.

FIGURE 3.11 – Fonction de Survie de la Perte d'emploi selon le Sexe

Les conclusions restent les mêmes que celles obtenues avec le modèle de Kaplan Meier. La population masculine a une fonction de survie plus faible que la population féminine.

Le type de produit

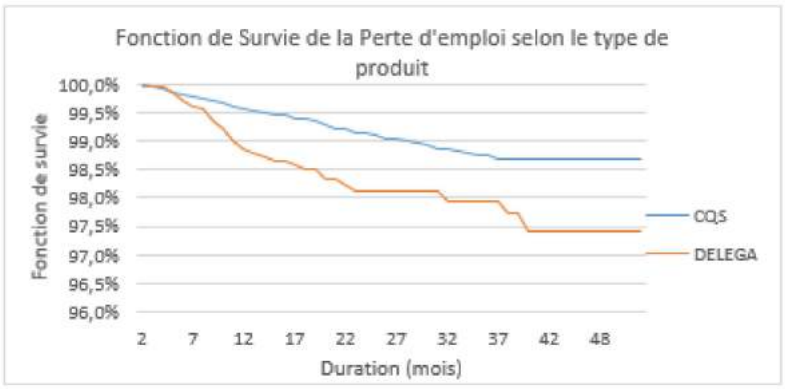


FIGURE 3.12 – Fonction de Survie de la Perte d'emploi selon le Type de Produit

Globalement, les conclusions sont les mêmes que celles obtenues au paragraphe précédent. Le produit Delega est plus risqué que le CQS.
Alors que les courbes obtenues avec la méthode de Kaplan Meier étaient très proches l'une de l'autre, la méthode de Cox illustre une différence plus significative entre celles-ci.

Fonctions de survie par individu

Finalement, le modèle de Cox nous permet d'obtenir les fonctions de survie de chacun des individus. Celles-ci peuvent nous aider à mieux prévoir et anticiper le risque et développer des stratégies adaptées.

A titre indicatif, nous choisissons de retenir les courbes des dix individus les plus à risques et des dix individus les moins à risques.

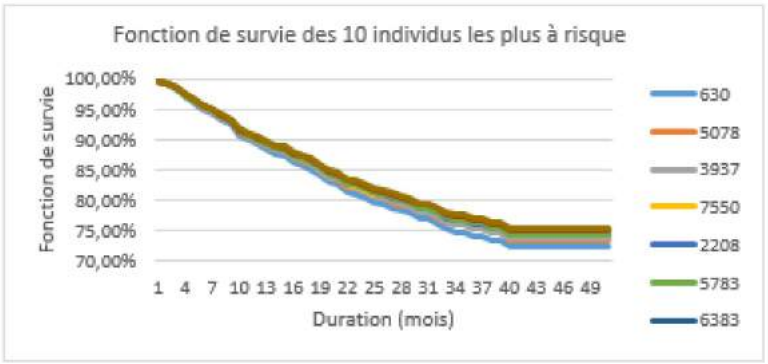


FIGURE 3.13 – Fonction de survie des 10 individus les plus à risque

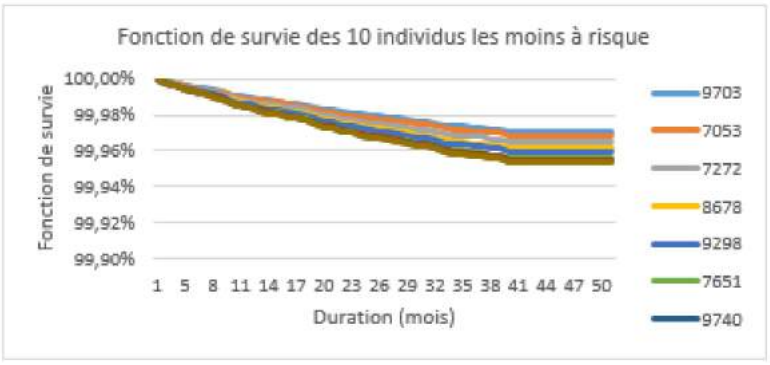


FIGURE 3.14 – Fonction de survie des 10 individus les moins à risque

Nous obtenons, en accord avec nos résultats précédent, en analysant les profils des individus sélectionnés, les conclusions suivantes : Les individus les plus à risques ont les caractéristiques suivantes :

- Ils sont employés du secteur privé
- Ils ont emprunté un Déléga
- Ils sont de sexe masculin
- Ils ont été financés par le Partenaire1
- Ils sont d'âges très variables
- Ils ont emprunté des montants très différents sur des durées variables

Les individus les moins à risques ont les caractéristiques suivantes :

- Ils sont employés du secteur public
- Ils ont emprunté un CQS
- Ils sont de sexe féminin
- Ils ont été financés par un autre partenaire
- Ils sont d'âges très variables
- Ils ont emprunté des montants très différents sur des durées variables

Nous avons exploité dans cette partie deux grandes méthodes d'analyses de survie : les modèles de Kaplan Meier et de Cox. Très adaptée à nos données censurées, elles s'appuyent notamment sur

l'étude des courbes de survie et permettent le calcul des taux d'incidence selon différentes modalité et la détermination de l'influence des variables explicatives.
En raison du faible nombre d'observations, les taux obtenus sont irréguliers, nous choisissons alors la méthode de lissage de Whittaker Anderson, la plus adaptée à nos données.

Après une description technique du fonctionnement des modèles, leur application permet de retenir plusieurs éléments.
Comme nous l'avions remarqué à la p précédente, la catégorie socio-professionelle est la variable la plus discriminante. Le taux de perte d'emploi est 10 fois plus élévé pour les employés du secteur privé que les fonctionnaires du public. Ceci s'explique par la précarité des emplois du secteur privé en comparaison avec les fonctionnaires qui bénéficient d'une sécurité d'emploi.
Sur notre portefeuille, les hommes sont plus sujets à la perte d'emploi que les femmes. Nous nous questionnons sur cette constation qui ne semble pas en ligne avec les taux de chômage en Italie (11% des femmes et 9% des hommes). Nous nous intérogeons donc sur la répartition de la population au sein de notre portefeuille.
Enfin, l'emprunt secondaire *Delega* en plus du CQS, augmente le risque d'insolvabilité.

Chapitre 4

Machine Learning

Au cours de ces dernières années les techniques d'apprentissage, grâce à la capacité à modéliser les relations et à la qualité de leurs prédictions globales, ont obtenu des succès significatifs. En analyse de survie, pour la plupart des applications du monde réel, l'objectif principal est d'obtenir une meilleure estimation du moment où se produit l'événement d'intérêt. L'un des principaux défis de ces études est la présence de censures, ainsi l'événement n'est pas observé en raison de la limitation de la période d'étude par exemple, et, l'information n'est que partiellement disponible sur certains individus. Par conséquent, il n'est pas approprié d'appliquer directement des algorithmes prédictifs utilisant les approches statistiques automatique standards.
C'est dans ce cadre que les approches de machine Learning ont été largement abordées dans la littérature pour apporter des solutions aux problèmes d'analyse de survie avec des algorithmes d'apprentissage adaptés aux particularités citées.

Nous présenterons dans cette étude les principales méthodes d'apprentissage automatique utilisées dans l'analyse de survie. Nous en ferons ensuite l'application.

4.1 Généralités sur le Machine Learning

4.1.1 Première approche

Le principe du machine Learning, branche de l'intelligence artificielle, est d'apprendre à l'ordinateur à partir des flux de données. L'idée est d'améliorer les performances des ordinateurs à résoudre un ensemble de tâches, sans être directement programmés pour chacune d'entre elles. Le principe de base de l'apprentissage automatique est de créer des algorithmes capables de recevoir des données d'entrées et d'utiliser une analyse statistique pour prédire une sortie tout en se mettant à jour au fur et à mesure que de nouvelles données sont disponibles. Lors de ce processus deux phases s'enchaînent : une phase d'apprentissage et une phase d'application. La littérature décrit aujourd'hui de nombreux modèles d'apprentissage automatique.

La première phase consiste à estimer un modèle à partir des données fournies à l'ordinateur. Ce processus s'appelle l'estimation d'un modèle, il consiste en la résolution d'une tâche telle que la reconnaissance d'un élément dans une image, la conduite autonome d'un véhicule... d'autres utilités existent. La seconde phase consiste à soumettre de nouvelles données au modèle dorénavant déterminé afin de résoudre la tâche souhaitée.

Selon les données fournies, l'apprentissage automatique se divise en deux catégories. L'apprentissage supervisé ou classification et l'apprentissage non supervisé ou clustering.
La principale différence entre les deux types d'apprentissage réside dans le fait qu'en apprentissage supervisé nous avons une connaissance préalable de ce que doivent être les valeurs en sortie de nos échantillons.

L'apprentissage non supervisé ne dispose pas de résultats étiquetés à l'avance. L'objectif est donc de déduire la structure naturelle inhérente aux données dont nous n'avons pas connaissance a priori. En général, des systèmes d'apprentissage non supervisés permettent d'exécuter des tâches plus complexes que les systèmes d'apprentissage supervisé, mais peuvent se montrer plus imprévisibles. Cette catégorie comprend une autre méthode de positionnement des données : "les K moyennes" (*K-means clustering*), ou encore, une autre possibilité de capter les données, la réduction de dimension (*Dimensionality réduction*) et les réseaux de neurones (*neural networks*)...

L'apprentissage supervisé a pour objectif d'apprendre une fonction qui, à partir d'un échantillon de données fournies et de résultats, se rapproche le plus de la relation entre les données d'entrées et de sorties observées. L'idée étant de repérer une structure ou relation particulière entre les données pour prédire ce qui est attendu. Cette catégorie comprend entre autres les arbres de décisions (*Decision Tree*), le Boosting, les forêts aléatoires (Random Forest) ... Lors de notre démarche d'étude et d'analyse, nous nous intéresserons uniquement à ce type d'apprentissage : en effet, dans notre base de données, nous avons déjà connaissance de la population des assurés qui ont perdu leur emploi et donc les résultats attendus.
Nous distinguons, ensuite, parmi l'apprentissage supervisé, les problèmes de régression des problèmes de classement. Les méthodes de régression permettent la prédiction d'une variable qualitative tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification. Le principal objet d'étude de ce mémoire étant la prédiction de la perte d'emploi à la suite d'une démission ou un licenciement, nous nous intéresserons davantage à la classification en deux classes : la survenance ou non du chômage.

En plus de présenter une grande variété de modèles, le machine Learning permet d'avoir une vision de la population par individu par individu. L'avantage de cette démarche est donc d'identifier individuellement les assurés faisant partie d'une population à risque et de distinguer les facteurs de risques les plus influents.

Nous pouvons ajouter que notre démarche est facilitée par l'accès aux données, la plupart des méthodes font peu d'hypothèses sur les données sous-jacentes au modèle.

4.1.2 La démarche de prédiction par Machine Learning

Afin de prédire l'incidence du chômage dans notre portefeuille via le Machine Learning, la démarche est classique à tout projet de machine Learning. Après s'être assuré de la qualité de nos données et avoir analysé avec des premières statistiques descriptives les variables explicatives issues de notre base d'apprentissage, nous effectuons un découpage de la base en deux parties, cette étape est appelée l'échantillonnage.
Ensuite, les modèles d'apprentissage seront appliqués sur l'échantillon d'apprentissage, appelé *train*, ainsi le modèle sortant de ce processus sera testé sur l'échantillon d'entraînement appelé *test*. Enfin, il faudra analyser les résultats obtenus et notamment comparer la robustesse et efficacité des modèles. Nous effectuerons ces étapes avec le language de programmation Python.

L'importance de l'échantillonnage

Cette étape est primordiale, elle consiste à répartir et diviser la base de données en deux échantillons, le *train* et le *test*. Normalement, la répartition est de 70% pour la base d'apprentissage et 30% pour la base de test. Nous utilisons la bibliothèque scikit-learn de Python, bibliothèque libre destinée à l'apprentissage automatique. Nous avons vu qu'un bon modèle de machine Learning doit être capable de faire des prédictions et de généraliser, nous permettant de tirer des conclusions dans le cadre de notre étude. Or si nous entraînons un modèle avec certaines données, il sera naturellement plus performant sur ces données que sur de nouvelles données et nous n'aurons donc une vision biaisée de ses capacités de prédiction. C'est pourquoi cette étape est primordiale, car, c'est en expérimentant le modèle sur la base *test* que nous connaîtrons ses performances.

Deux phases : l'apprentissage et la prédiction

Les phases d'apprentissage et de prédiction reposent sur les algorithmes choisis en fonction des variables explicatives, de la taille de la base de données, du temps de calcul ou de la précision nécessaire.

A ce stade, nous utilisons la validation croisée, qui consiste en une utilisation de toutes les données de la base à la fois pour l'entraînement et pour la validation. Nous partageons les données en k parties égales, appelés k-folds. A chacune des étapes nous utilisons k-1 parties pour entraîner le modèle et la dernière pour le test. Nous réitérons l'opération avec, à chaque fois, une des k partie différente. Finalement, chaque élément de la base sera utilisé une fois comme test et k-1 fois comme donnée d'entraînement.
Nous évaluons finalement la performance du modèle, moyennant celles obtenues pour chaque essai.
Il est important de stratifier la validation : lors du partage des données, des classes équilibrées doivent être formées dans le sens ou chacune d'entre elle doit contenir environ le même nombre d'éléments de chacune des classes afin d'éviter de biaiser les résultats.

En cas de déséquilibre des classes nous effectuons un ré-échantillonnage.

Par convention, nous choisissons, pour notre étude $k = 10$.

3) Une démarche qui contourne Lles limites des algorithmes

Le Machine Learning reste une science très complexe, notamment car il n'existe pas de modèle parfait qui puisse correspondre à toutes les situations. En effet, pour chaque problème posé, il est essentiel d'analyser et nettoyer minutieusement les données, formuler des hypothèses et surtout tester plusieurs modèles et algorithmes afin de déterminer le plus pertinent dans notre situation.

Nous sommes souvent confrontés à des modèles très complexes et coûteux en temps ce qui limite leur utilisation. Pour éviter cela, et simplifier la convergence des algorithmes, nous cherchons à simplifier au maximum les problèmes en minimisant la perte d'informations. En effet nous choisirons donc un nombre limité de variables d'entrées. Il peut aussi être nécessaire de réduire la taille de la base de données fournie. Dans notre cas, la base de données étant justement très petite, nous ne modifierons pas sa taille. Enfin, nous pouvons réduire la complexité des hypothèses de modélisation.

Nous introduisons alors le compromis biais variance.
Ce-dernier vise à diminuer les sources d'erreurs qui empêchent la généralisation des algorithmes de

Machine Learning au-delà de l'échantillon d'apprentissage. Idéalement, nous cherchons un modèle qui reflète les régularités des données d'apprentissage mais qui peut aussi s'étendre aux données test.

Le biais correspond à une erreur provenant des hypothèses utilisées. Un biais élevé peut être lié à un sous apprentissage. La variance correspond à une erreur provenant de la sensibilité aux petites variations de l'échantillon d'apprentissage.

Nous savons que le sur-apprentissage (*overfitting*) se produit lorsqu'un modèle est excessivement complexe, ainsi il se rapproche des données jusqu'à prendre en compte le bruit, que nous souhaitons naturellement exculre de notre étude. Ainsi, cela se produit souvent lorsqu'il y a trop de variables d'entrées, et dans ce cas, il est alors probable que le modèle corresponde parfaitement aux données d'entraînement. Cette analyse pourrait nous sembler satisfaisante mais elle ne correspond pas aux effets escomptés. En effet, bien que le modèle corresponde aux données d'entraînement, il aura tendance à échouer à prédire des valeurs nouvelles et donc à généraliser. Cet évènement apparait lorsqu'il y a une dépendance sur les données et les fluctuations associées notamment, qui ne s'appliquent pas nécessairement à des nouvelles données. Nous verrons que ce dysfonctionnement se traduit par une variance trop élevée de la prévision.

A contrario le sous-apprentissage (*underfitting*) se traduit par un biais du modèle trop élevé. Dans ce cas, le modèle ne se rapproche pas suffisamment des données et ne parvient pas à en capter la tendance sous-jacente. Le modèle est alors inadapté et ne pourra pas généraliser les prédictions. Cet effet peut s'expliquer par un nombre insuffisant de données ou de variables.
Ces différents phénomènes sont résumés sur le schéma ci-dessous :

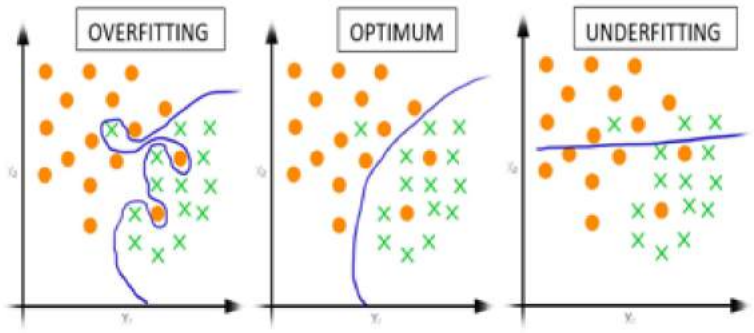


FIGURE 4.1 – Schéma de l'overfitting et de l'underfitting

Nous constatons que la décomposition biais-variance permet d'évaluer l'espérance de l'erreur de prédiction d'un algorithme d'apprentissage comme la somme de trois termes : le biais, la variance et l'erreur irréductible.
Nous disposons par exemple de nos données d'apprentissage composées de n points $(x_1, y_1),\ldots,(x_n, y_n)$. Nous supposons qu'il existe une fonction f telle que :

$$\mathrm{E}\left[(y - \hat{f}(x))^2\right] = \mathrm{Biais}[\hat{f}(x)]^2 + \mathrm{Var}[\hat{f}(x)] + \sigma^2 \text{ avec :}$$

— $\mathrm{Biais}\ [\hat{f}(x)] = \mathrm{E}[\hat{f}(x) - f(x)]$
— $\mathrm{Var}[\hat{f}(x)] = \mathrm{E}\left[(\hat{f}(x) - \mathrm{E}(\hat{f}(x)))^2\right]$

Nous cherchons donc une fonction telle que :
— Le biais peut être compris comme l'erreur due aux hypothèses simplifiées de la méthode
— La variance correspond au déplacement de la fonction $\hat{f}$ autour de sa moyenne
— L'erreur irréductible résultant du bruit dans le problème lui-même

La complexité du modèle entraîne un biais faible. Mais cette complexité rend le modèle davantage mobile autour de la moyenne, donc la variance augmente.
En cas de variance trop élevée nous pouvons réduire le nombre de dimensions du modèle, c'est-à-dire le nombre de variables afin de simplifier le modèle.

Comparaison des modèles

Il sera nécessaire de comparer les modèles afin de finalement choisir le plus performant et le plus adapté à notre base de données. Nous cherchons le modèle dont la complexité est optimale. Il existe plusieurs indicateurs et méthodes de sélection que nous analyserons plus en détail dans les parties suivantes.

4.2 Les algorithmes de machine Learning

Dans cette partie nous présenterons les algorithmes de machine Learning les plus utilisés en analyse de survie, auxquels nous nous restreindrons dans ce mémoire. Il existe en effet une grande variété d'autres algorithmes d'apprentissage, mais qui ne s'appliquent en l'espèce à nos données.

Nous choisissons d'implémenter trois modèles de classification qui sont :
— le CART
— les forêts aléatoires
— le Gradient Boosting

L'objectif de cette partie est de décrire et expliciter le cadre théorique de ces quelques algorithmes qui seront utilisés pour prédire l'incidence de la perte d'emploi.

Il existe de nombreux autres algorithmes sur lesquels nous avons choisi de ne pas nous attarder, soit parcequ'ils ne sont pas adaptés à nos variables ou bien les tests qui nous sont restitués montrent des résultats non concluant et de mauvaise qualité.

Comme nous l'avons vu précédemment, lors de l'analyse de survie, le modèle de Cox est souvent utilisé en raison de sa practicité. Ainsi, des extensions d'approches de machine learning ont permis d'obtenir aussi de bons résultats, qu'il nous paraît opportun de comparer avec les prédictions de la régression de Cox.

4.2.1 Les arbres de décision : CART

Les arbres de décision permettent de résoudre des problèmes de classification et de régression. Les arbres de régression permettent de prédire une valeur numérique tandis que des arbres de classification permettent de prédire la classe à laquelle appartient une variable de sortie. Ce sont des modèles simples qui permettent d'obtenir des résultats intuitifs et facilement compréhensibles, représentés sous forme graphique d'un arbre. Ils sont construits suivant des règles de classification basées sur des tests associés aux variables et organisés en arborescence. L'algorithme de CART (*Classification And Regression*

Trees) a été développé par Breiman, Friedman, Olshen et Stone en 1984. Il permet un développement important des arbres de décision. Il définit une partition de l'espace des variables d'entrées, puis, sur chaque partition, un modèle simple est ajusté. La partition se fait via des conditions binaires.
Nous rappelons qu'un arbre de décision est une structure hiérarchisée formée de nœuds. Le nœud initial est la racine. Chaque nœud se divise en au maximum en deux nœuds, un droit et un gauche.

Toutes les branches qui arrivent à un nœud correspondent à toutes les valeurs de la variable d'entrée telle que les arrêtes arrivant à un nœud couvrent l'ensemble des valeurs possibles. Ainsi, à chacun des nœuds, une variable d'entrée est choisie et répartie en deux groupes. A chaque étape, les données sont séparées en deux sous ensemble en fonction de la valeur d'entrée. Ce processus est répété sur chaque sous ensemble récursif de manière itérative.

Chaque nœud n'aillant aucun fils (descendant) est une feuille. Les feuilles correspondent aux variables cibles et les embranchements à des combinaisons de variables d'entrées qui amènent à ces valeurs.

Au niveau des feuilles, nous représentons la valeur de la variable cible ou bien une distribution de probabilité des différentes valeurs possibles de cette même variable cible.

On dispose d'un échantillon d'apprentissage $\mathcal{L} = \left\{(X_i, Y_i)_{i\in\{1,\dots,n\}}\right\}$ où :
- $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ sont n réalisations indépendantes d'un couple de variables $(X \cdot Y) \in \mathcal{X} \times \mathcal{Y}$
- $\forall i \in \{1,\dots,n\}, X_i = \left(X_i^1 \dots X_i^p\right)$ vecteur de taille p contenant les observations pour l'individu i des p variables explicatives
- Y_i est l'observation pour l'individu i de la variable à expliquer

Nous sommes dans le cadre d'une classification : $\mathcal{Y} \subset \mathcal{J}$ et $Y = s(X)$.
Nous cherchons à estimer la fonction s.

L'algorithme de construction d'un arbre CART se compose de trois étapes :
- La construction de l'arbre maximale
- L'élagage
- La sélection finale

La construction de l'arbre maximal

Nous décrivons les étapes de l'algorithme.

Critères de construction

Cadre de la classification $\mathcal{J} = \{1, .., J\}$ Nous avons les notations suivantes :
- π_j la probabilité à priori de la classe j. Peut être estimée par $\frac{N_j}{n}$ avec $N_j = \text{Card}\{(x_k, y_k) \mid y_k = j\}$
- soit t un noeud de l'arbre, $N_t = \text{Card}\{(x_k, y_k) \mid x_k \in t\}$ nombre d'observations de $\mathcal{L}$ dans t
- soit t un noeud de l'arbre et $j \in \mathcal{J}$, $N_j = \text{Card}\{(x_k, y_k) \mid x_k \in t \text{ et } y_k = j\}$ nombre d'observations de $\mathcal{L}$ dans t et de classe j
- $P(j, t)$: probabilité qu'une observation soit dans le noeud t et de classe j
  $\Rightarrow$ estimée par $p(j, t) = \pi_j \frac{N_j(t)}{N_j}$
- $P(t)$: probabilité qu'une observation soit dans le noeud t
  $\Rightarrow$ estimée par $p(t) = \sum_j p(j, t)$

— $P(j|t)$: probabilité a posteriori dans t de la classe j
$\Rightarrow$ estimée par $\frac{p(j,t)}{p(t)}$

Nous souhaitons répartir les individus en groupes plus homogène, au sens de notre variable d'intérêt, et nous définissons donc la fonction d'hétérogénéité.

Définition Soit h une fonction de $\left\{(p_1, \ldots, p_j) \mid p_i \geq 0, \sum_j p_j = 1\right\}$ dans $\mathbb{R}$. h est une fonction dite d'hétérogénéité si :

— h est symétrique en $p_1, ..., p_J$
— h est maximale en $(\frac{1}{J}, .., \frac{1}{J}$
— h est minimale en $(1, 0, ..., 0), ..., (0, ..., 0, 1)$

Définition Soit t un noeud. On définit l'hétérogénéité de t par :
$i(t) = h(p(1|t), ..., p(J|t))$ avec h une fonction d'hétérogénéité.

Dans le cas des algorithmes CART nous utilisons usuellement la fonction d'hétérogénéité de Gini :
— Gini $h(p_1, ..., p_J) = \sum_{i \neq j} p_i p_j$
— Shanon $h(p_1, ..., p_J) = -\sum_j p_j log(p_j)$

Nous avons la possibilité d'utiliser la fonction de Shannon. Cette fonction associe une hétérogénéité à chaque nœud. Elle est de 0 pour un nœud qui ne possède que des observations d'une même classe. Elle est maximale si un nœud possède autant d'individus dans chacune des deux classes. Alors nous pouvons comprendre la fonction d'hétérogénéité de Gini comme la probabilité d'obtenir deux observations de classes différentes.

Principe de construction d'une division Soit t un noeud de l'arbre. Soit t_d son descendant de droite et t_g son descendant de gauche, descendants engendrés par une division δ.
On note $p_g = \frac{p(t_g)}{p(t)}$ et $p_d = \frac{p(t_d)}{p(t)}$: proportion d'observations envoyées respectivement dans t_g et t_d.

La variation d'hétérogénéité générée par δ est définie par :

$$\Delta i(\delta, t) = i(t) - p_g i\,(t_g) - p_d i\,(t_d)$$

La division optimale du noeud t est donnée par :

$$\delta^*(t) := \delta^* = \underset{\delta \text{ division}}{\operatorname{argmax}} \Delta i(\delta.t)$$

La construction ne sera valide que si h est une fonction concave c'est à dire : $\Delta i(\delta, t) \geq 0$

Règle d'arrêt Par récursivité, de la procédure décrite à la partie précédente la construction de l'arbre est alors évidente.
Un noeud t d'un arbre est déclaré terminal si :
— il n'y a qu'une seule observation dans le noeud t
— il n'y a, dans t, que des observations avec un même label

Règle d'assignation

définition Soit t un noeud dont la réponse associée est $j(t)$.
La probabilité de mauvais classement du noeud t évaluée par substitution, est définie par : $r(t) = \sum_{j \neq j(t)} p(j \mid t)$

définition Soit t un noeud.
La réponse $j(t)$ associée est définie par : $j(t) = \underset{j \in J}{\text{argmax}} p(j \mid t)$

Propriété Cette définition de $j(t)$ est telle que $r(t)$ est minimale.

Nous affectons chaque feuille à une des valeurs de la variable à expliquer. Dans le cas d'une variable d'intérêt quantitative, nous associons à chaque feuille la moyenne des observations relatives à cette feuille. Dans le cas d'une variable qualitative, nous associons chaque feuille à la modalité la mieux représentée dans le nœud.

L'élagage de l'arbre

L'élagage de l'arbre a pour objectif de construire des arbres plus raffinés et pallier la complexité de l'arbre maximal. Ce dernier est souvent trop grand et sujet au sur-apprentissage, dont nous avons précédemment présenté les dangers. A chaque étape des sous arbres sont construit, dérivés de l'arbre initial, pour ensuite définir l'arbre optimal.

Notations
- Soit T un arbre et t un noeud non terminal de T. Elaguer T à partir de t consiste à créer un nouvel arbre T^* qui n'est autre que T privé de tous les descendants de t.
- Tout arbe $T^{'}$ obtenu par élagage de T est un sous arbre de T ce que l'on note $T' \prec T$.

Soit un arbre T construit, avec k feuilles, nous pouvons mesurer la qualité de discrimination de cet arbre par le critère : $D(T) = \sum_{i=1}^{k} D_i(T)$ où $D_i(T)$ désigne le nombre de mal classés. Les sous-arbres sont alors construits pas à pas en pénalisant la complexité de l'arbre. Celle-ci est définie par $C(T) = D(T) + \gamma k$. Nous faisons varier γ pour obtenir des sous arbres imbriqués les uns dans les autres et notamment imbriqués dans l'arbre maximal. Ainsi la complexité est un paramètre compris entre 0 et 1 qui détermine la taille de l'arbre. Plus la complexité est faible, plus l'arbre est complexe et donc risque le sur-apprentissage.

Ainsi, l'élagage permet de réduire le nombre de feuilles totales de l'arbre, en utilisant le critère de pénalité permettant d'obtenir le meilleur compromis entre la précision et la complexité.

La sélection finale

Nous disposons, après cette phase d'élagage de plusieurs sous arbres et donc de plusieurs estimateurs. Nous souhaitons sélectionner le plus adapté. Pour cela, nous disposons d'un échantillon test. L'arbre final retenu est celui avec la plus faible erreur estimée sur les données test.

Les arbres de survie

Nous utiliserons les arbres de survie. Ils sont une forme d'arbres de décisions adaptés pour traiter les données d'analyse de survie. Nous rappelons en effet que les données sont censurées.

Le principe des arbres est de partitionner récursivement les données en fonction d'un critère de fractionnement particulier et les individus similaires sont alors placés dans la même branche. Son objectif est de prédire la valeur de la variable cible en fonction de plusieurs variables d'entrée.

Ainsi, la principale différence entre un arbre de survie et l'arbre de décision standard réside dans le choix du critère de fractionnement. L'arbre de décision se construit en une partition récursive des données en fixant un seuil pour chaque variable, mais elle ne peut pas rendre en compte les interactions entre les variables ni les informations censurées dans le modèle. Ainsi l'arbre de décision a la capacité de traiter les données censurées en utilisant la structure arborescente.

Les critères utilisés pour les arbres de survie peuvent être regroupés en deux catégories :
— Maximiser l'hétérogénéité entre les nœuds
— Minimiser l'homogénéité à l'intérieur des nœuds
Nous utilisons donc un critère approprié pour les données censurées à droite, le test du log rank, le test le plus populaire pour comparer des courbes de survie. C'est un test non-paramétrique permettant de considérer toute l'information sur l'ensemble du suivi, sans la nécessité de faire des hypothèses sur la distribution des temps de survie.

Les limites des arbres de décision

Les arbres de décision permettent une lecture et une interprétation très simple. Néanmoins, ils présentent de nombreuses limites.

Les arbres de décisions sont souvent sujets au risque de sur-apprentissage. Ceci signifie que le modèle apprend très précisément les données, à tel point qu'il s'y ajuste complètement et enregistre les bruits associés. Les données utilisées pour l'apprentissage sont ainsi reconnues cependant le modèle est incapable de prédire et de s'adapter à de nouvelles données. Or une des capacités essentielle pour un modèle de Machine Learning réside dans la capacité à apprendre des données pour ensuite prédire de nouvelles données et anticiper des risques et leurs impacts. L'élagage tend à réduire cet effet de sur-apprentissage mais reste insuffisant. De ce fait, les arbres construits sont très instables et trop proches des données. Une légère variation de celles-ci entraîne une modification importante de la structure de l'arbre et des résultats obtenus.

4.2.2 Les méthodes d'agrégation

Nous nous intéressons dans cette partie aux méthodes d'agrégation de *Machine Learning*.

Le Bagging - Les forêts aléatoires

L'algorithme des forêts aléatoires (ou *random forest*) a été développé par Breiman en 2001. C'est un algorithme de classification, permettant de réduire la variance des prévisions d'un seul arbre de décision. Cet algorithme effectue un apprentissage sur plusieurs arbres de décisions construits

aléatoirement et entraînés sur des sous-ensemble de données légèrement différents les uns des autres. Pour cela, il effectue une approche de type *bagging*.

Le *Bagging* (ou *Boostrap aggregation*) est une technique permettant d'améliorer la classification des arbres de décision. Il permet de réduire la variance associée à l'estimateur et donc améliorer la stabilité et performance des arbres de décision.

Soit Y la variable à expliquer, $X_1, \ldots, X_p$ les p variables explicatives, ϕ le modèle appris sur un échantillon aléatoire $Z = \{(x_1, y_1), ...(x_n, y_n)\}$ où x_i sont décrit par p variables explicatives. En effectuant B tirages aléatoires avec remise sur Z on obtient B échantillons bootstrap $z_1, \ldots, z_B$ parmi Z. Sur chaque échantillon est appliqué un modèle ϕ_{z_i}. La prédiction est ensuite réalisée en agrégeant les décisions de chaque modèle en effectuant un vote par majorité dans le cas d'une variable qualitative, ou une moyenne dans le cas d'une variable quantitative.
L'erreur est ensuite calculée par une estimation « out-of-bag ». Pour chaque observation (x_i, y_i), l'erreur *out of Bag* (OOB) est l'erreur moyenne calculée en utilisant les prédictions des arbres qui ne contiennent pas cette observation dans leurs échantillons Boostrap respectifs, sur lesquels ils ont été entraînés. L'erreur OOB décroit avec le nombre de modèles utilisés puis se stabilise lorsque le nombre optimal de modèles à utiliser est atteint.

Pour les Forêts aléatoires nous disposons donc d'un échantillon d'apprentissage $Z = \{(x_1, y_1), ..., (x_n, y_n)\}$. $X_1, \ldots, X_p$ sont les p variables explicatives. En construisant B arbres aléatoires dans la forêt, le Bagging effectue B tirages aléatoires avec remise parmi Z. L'algorithme crée ensuite un arbre sur chaque échantillon, en réalisant, pour la construction de chaque noeud un tirage de m variables parmi les p variables disponibles afin former la décision associée au noeud.
La prédiction correspond alors à la classe majoritaire.

Le nombre optimum d'arbre, pouvant être très nombreux, est un paramètre important, fonction des données et variables du problème. Nous l'obtenons en analysant l'erreur OOB en fonction du nombre d'arbres utilisés. Elle se stabilise lorsque le nombre optimal d'arbres à utiliser est atteint.

Nous avons présenté précédemment les difficultés liées au sur-apprentissage. Les forêts sont une alternative permettant d'éviter ce problème, en effet, elle procède suivant la majorité, ce qui diminue la marge d'erreur d'un arbre seul, qui pose un problème d'optimalité. Comme nous l'avons exposé dans la partie précédente, ceux-ci sont construit par étape, en sélectionnant, à chaque étape le test le plus judicieux. Or, chaque test n'apparaissant qu'une seule fois dans l'arbre, les résultats obtenus peuvent être loin des résultats attendus.

Généralement, il s'agit d'une des méthodes les plus efficace et les plus utilisés.

Dans le cadre de l'analyse de survie, nous utilisons les forêts aléatoires de survie (ou *Random Survival Forest*), qui sont une extension des forêts aléatoires classiques, tenant compte des censures.

Les forêts sont formées en combinant les résultats de nombreux arbres de survie. De nouvelles règles de fractionnement pour les arbres de survie sont introduites, précisées précedemment.

Les forêts présente l'avantage de contourner la nécessité d'imposer des contraintes paramétriques ou semi-paramétriques sur les distributions sous-jacentes et permettent des prédictions précises.

La méthode générale est la même, néanmoins, au moment de la division des nœuds des arbres en

deux sous-ensembles, nous utilisons un critère approprié pour les données censurées, le test du log rank.

Le Boosting - XG Boost

Le gradient Boosting est une méthode de Machine learning pour les problèmes de classification et de régression. Il produit un modèle de prédiction à partir d'un ensemble de modèles de prédictions faibles, comme des arbres de décision. Les algorithmes de Boosting sont des algorithmes itératifs de descente de gradients fonctionnels. Ils ont pour but de trouver les valeurs optimales des paramètres d'une fonction donnée. Il cherche à ajuster les paramètres afin de minimiser la fonction de perte du modèle en utilisant une descente de gradient.

La descente de gradient est un algorithme d'optimisation itérative de premier ordre pour trouver un minimum local d'une fonction différentiable. Etant donné que le gradient Boosting est basé sur la minimisation d'une fonction de perte, différents types de fonctions de perte peuvent être utilisées, permettant une flexibilité de la méthode selon le problème posé.

L'idée se rapproche de celle du bagging qui, plutôt que d'utiliser un seul modèle, en utilise plusieurs, qui sont ensuite agrégés, afin d'obtenir un unique et meilleur résultat. Le Boosting construit le modèle par étape. Pour commencer, il construit un premier modèle qu'il évalue. Alors chaque individu est pondéré en fonction de la performance de la prédiction, affectant un poids plus important aux individus dont la valeur a été mal prédite pour la construction du modèle qui suit. Puis, à chaque étape, les poids sont ainsi corrigés, permettant de mieux prédire les valeurs moins faciles.

L'intérêt de cette méthode est de réduire la variance, tout comme le bagging, mais aussi le biais de la prévision.

Les limites des modèles d'agrégations

Bien qu'ils soient plus stables et qu'ils aient un meilleur pouvoir prédictif que les modèles simples, les modèles d'agrégation présentent néanmoins certaines limites.
Leurs temps de calcul peuvent être très long, bien plus élevés que pour les modèles simples. En effet, il s'agit de répéter de nombreuses fois un modèle simple jusqu'à atteindre une stabilisation de l'erreur de prévisions *out of bag* (erreur OOB).

Ces modèles requièrent des espaces de stockage suffisants pour stocker les différentes combinaisons du modèle et les utiliser ensuite sur de nouvelles données.

Enfin, le principal inconvénient est que, contrairement aux modèles simples comme les arbres de décisions qui sont très clairement et lisiblement interprétables, les modèles d'agrégation le sont beaucoup moins. Bien qu'ils améliorent la précision et la qualité de prédiction, ils sont bien plus complexes. L'une des seule information que nous pouvons en extraire est finalement l'importance des variables. Pour chacune des variables explicatives, nous pouvons déterminer son influence sur la variable cible, à savoir dans quelle sens elle influe, si elle augmente ou non l'exposition au risque et à quelle intensité. Ce sont des conclusions manifestement très intéressantes, car connaître l'importance des variables permet d'adapter la tarification des produits d'assurance en conséquence mais aussi de surveiller la souscription de ceux-ci et prêter une attention particulières à certains assurés répondant à des critères particuliers.

4.3 Evaluer les performances des modèles de machine learning

4.3.1 Qu'est-ce qu'un bon modèle ?

Nous avons présenté quelques algorithmes de Machine Learning, les principaux utilisés dans notre contexte, mais ils sont très nombreux. La plupart ont des hyperparamètres qui doivent être fixés. Un bon modèle de *machine Learning* doit permettre de généraliser, à savoir être capable de faire des prédictions à partir des données utilisées pour mettre en place le modèle mais surtout à partir de nouvelles données. C'est pour cette raison que nous appelons ces méthodes : l'apprentissage.

Dans le cas d'un modèle trop sensible aux variations des données, qui a de moins bonnes performances sur les nouvelles données de test, nous avons évoqué le sur-apprentissage. Ces modèles sont plus complexes que la réalité. Or, nous préférons la simplicité. Nous adoptons ainsi le principe d'Ockham selon lequel il n'est pas pertinent d'utiliser de nouvelles hypothèses tant que celles déjà énoncées suffisent, autrement dit, il n'est pas nécessaire d'apporter aux problèmes une autre réponse spécifique, avant de s'assurer que cela est indispensable. Cela complexifierai le problème et nous pousserai à passer à côté de la réalité recherchée.
D'autant plus qu'il existe des bruits autour des données. Celles-ci présentent en effet des erreurs de mesure ou d'étiquetage par exemple, c'est pourquoi les modèles ne doivent pas capter les bruits inhérents aux données.

A contrario, dans notre démarche, nous souhaitons aussi éviter les modèles trop simplistes qui ne capterons pas le phénomène recherché et ne permettront pas de faire des prédictions pertinentes, le sous-apprentissage.

4.3.2 Les critères de sélection

Plus spécifiquement, notre étude se place dans le cadre d'une classification binaire : les données dont nous disposons sont étiquetées afin de prédire leur appartenance ou non des données à une classe. Aussi, nous cherchons, à terme, l'appartenance ou non à la classe perte d'emploi.

Nous utilisons la matrice de confusion qui permet de déterminer la qualité d'un algorithme de classification. Les colonnes de la matrice correspondent au nombre d'occurrences d'une classe estimée tandis que les lignes correspondent au nombre d'occurrences d'une classe réelle.

La cellule ligne L, colonne C contient le nombre d'éléments de la classe réelle L qui ont été estimés comme appartenant à la classe C.

La matrice ci-dessous permet d'avoir une idée claire et rapide des capacités de classification d'un modèle. Elle indique, à partir des résultats escomptés et des prédictions, le nombre de prédictions correctes et incorrectes pour chaque classe, ce qui permet de connaître les erreurs commises et notamment le type d'erreur commis.

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

TABLE 4.1 – Matrice de confusion

Les terminologies de la matrice de confusions sont les suivantes :

— TP (*True Positive*) : les cas où la prédiction est positive, et où la valeur réelle est effectivement positive. Exemple : nous prédisons une survenance du chômage pour un individu, et il subit bel et bien le chômage avant la fin du contrat.

— TN (*True Negative*) : les cas où la prédiction est négative, et où la valeur réelle est effectivement négative. Exemple : nous ne prédisons pas de survenance du chômage pour un individu, et il ne subit effectivement pas de chômage avant la fin du contrat.

— FP (*False Positive*) : les cas où la prédiction est positive, mais où la valeur réelle est négative. Exemple : nous prédisons une survenance du chômage pour un individu, mais il ne subit pas le chômage avant la fin du contrat.

— FN (*False Negative*) : les cas où la prédiction est négative, mais où la valeur réelle est positive. Exemple : nous ne prédisons pas de survenance du chômage pour un individu, mais il subit le chômage avant la fin du contrat.

A l'aide de cette matrice de confusion, il est possible de calculer plusieurs critères que nous chercherons à optimiser, qui dénotent de la performance des modèles.

— Le rappel (ou *recall*) est le taux de vrais positifs

— La précision est la proportion de prédictions correctes parmi les points prédits positifs

— Le $f1$ score, qui est la moyenne harmonique du rappel et de la précision :
$F - mesure = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} = \frac{2TP}{2TP+FP+FN}$

4.4 Applications du modèle

4.4.1 Les limites de nos données

Nous rappelons que notre base de données comprend uniquement une population 2% d'assurés sinistrés, soit 170. En répartissant nos données de cette population en deux classes, les sinistrés et les non sinistrés nous obtenons alors des classes déséquilibrées. Celles-ci vont poser un problème au niveau de l'*Accuracy* (la précision) lors des modélisations de Machine Learning.
Nous notons que la précision standard ne mesure plus de façon fiable le rendement, ce qui rend la formation du modèle beaucoup plus difficile. Pour les enregistrements qui n'ont pas eu d'entrée au chômage, la précision sera de 100%, pour la population qui a perdu son emploi elle sera de 0%. La précision globale sera très élevée simplement parce que la plupart de la population des assurés n'a pas été concernée par une entrée en chômage (et non pas parce que le modèle est correct).

Ce dysfonctionnement nous est dommageable, car de nombreux algorithmes d'apprentissage automatique sont conçus pour maximiser la précision globale. Nous allons donc utiliser une technique pour la gestion des classes déséquilibrées.

Donc, pour remédier à ce déséquilibre, nous allons utiliser l'échantillonnage ascendant, un processus qui consiste à reproduire au hasard les observations de la classe minoritaire afin de renforcer son signal. Plusieurs heuristiques nous permettent de le faire mais la méthode la plus courante, que nous allons

utiliser, est de simplement rééchantillonner selon le remplacement. Nous importons donc le module de rééchantillonnage de Sckit-Learn.

Nous allons donc créer une nouvelle base de données avec une classe minoritaire échantillonnée :

— Premièrement nous séparons les observations de chaque classe selon l'information sur la sinistralité (en différentes bases de données).
— Ensuite, nous rééchantillonnons la classe minoritaire avec remplacement, en fixant le nombre d'échantillons à 3300 unités pour correspondre davantage à celui de la classe majoritaire.
— Enfin, nous combinerons la classe minoritaire échantillonnée avec la classe majoritaire, obtenant ainsi une nouvelle base de données composée à 25% de contrats ayant subi un sinistre et les 75% restant ne subissent pas de sinistres.

Désormais les modèles obtenus, bien que de précisions plus faibles, seront plus performants et les résultats plus significatifs.

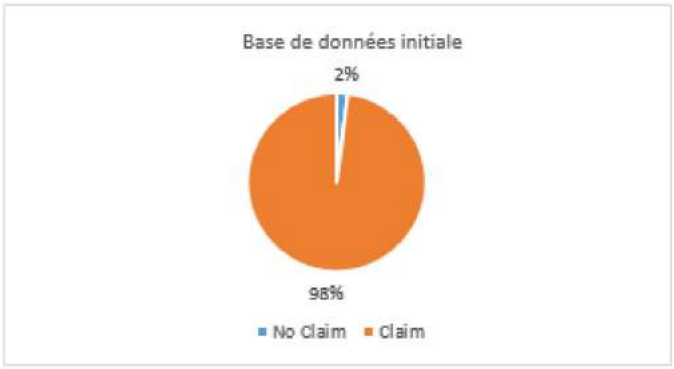


FIGURE 4.2 – Répartition des données dans la base de données initiale

FIGURE 4.3 – Répartition des données dans la base de données rééchantillonnée

Etant donné que les prévisions faites par les algorithmes de *Machine Learning* sont basées uniquement sur les données fournies à l'ordinateur, il est primordial d'en faire au préalable un traitement et un nettoyage très minutieux. Cette phase d'analyse, de nettoyage et de correction a été préalablement effectuée sur notre base de données et nous l'avons évoquée au chapitre 2.

4.4.2 Les arbres de survie

Les arbres de survie permettent de répartir la population en classes, selon les caractéristiques des individus. Cette répartition peut permettre de tarifier les produits d'assurance en fonction des caractéristiques associées à chaque individu de la population assuré.

Le graphique ci-dessous nous permet de constater qu'au niveau de chaque nœud, nous disposons de plusieurs informations. Ainsi, par exemple, pour le premier nœud :

- Socio_professional_category_PRIV ≤ 0.5 : cela correspond à la variable et la valeur qui divise le nœud en deux
- L'information Gini $= 0,361$: il s'agit en l'espèce de l'indice Gini du nœud
- Sample $= 9044$: il s'agit du nombre de données dans ce nœud
- Value $= [6903; 2141]$: cette ligne indique l'estimation des individus, au niveau de ce nœud, dans chaque catégorie, dans l'ordre. Une population de 6 903 individus sont estimés dans la catégorie No LoE (*No Loss of employment i.e.* pas de Perte d'emploi) et 2 141 individus sont estimés dans la catégorie LoE (*Loss of Emplyment i.e.* perte d'emploi)
- Class = No LoE : cela correspond à la classe prédite, à l'étape actuelle, pour les éléments du nœud

Nous rappelons que le critère de division de l'arbre est le coefficient de Gini, mesure permettant de rendre compte de l'hétérogénéité des classes. Il vaut 0 pour un nœud qui ne possède que des observations d'une même classe et 1 pour un nœud qui possède le même nombre d'individus dans chacune des deux classes. La construction de l'arbre repose sur un critère de pénalisation qui complexifie de proche en proche l'arbre.

Nous avons tout d'abord obtenu un arbre, dont le degré de complexité était minime. Il possédait ainsi un très grand nombre de feuilles très raffinées. Puis, afin d'éviter le sur-ajustement excessif et nous avons procédé à un élagage de l'arbre. Cette opération nous a permis d'obtenir un arbre final plus parcimonieux et donc plus pertinent et précis.

Résultat obtenus La figure ci-dessous représente l'arbre de survie optimal sur notre portefeuille. Les valeurs obtenues dans les feuilles terminales indiquent les catégories d'assurés qui font parti de la population à risque. Les classes oranges de population ont une faible probabilité de perte d'emploi tandis que les classes bleues sont plus à risque.
Deux catégories de population sont identifiées comme à risque.

Nous trouvons ainsi :

- Les individus du secteur privé dont le montant assuré est supérieur à 19 000 euros et l'âge inférieur à 28 ans.
- Les individus du secteur public ayant effectué un prêt *Delega* dont la durée est inférieure à 30 mois.

Les autres emprunteurs ne semblent pas présenter de gros risques.

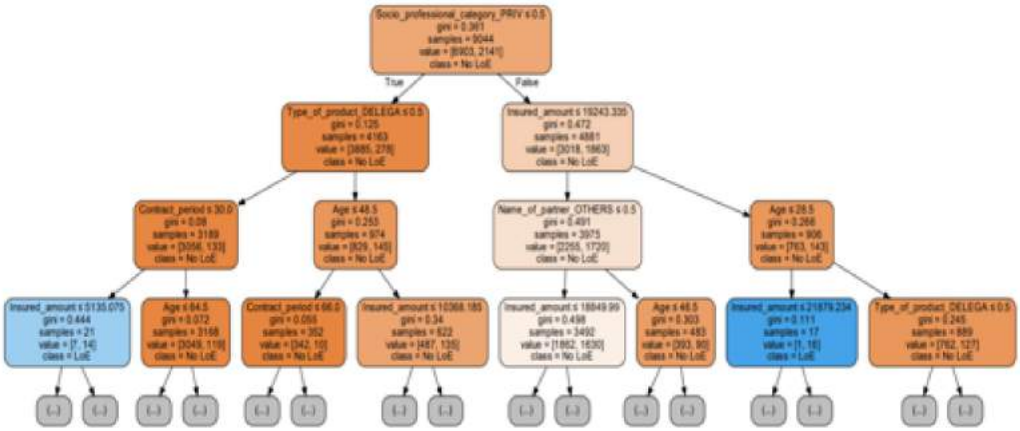


FIGURE 4.4 – Arbre de survie de notre portefeuille

Rappelons que cet arbre reste très instable et dispose de faibles capacités de prédiction. Ainsi nous préférerons les méthodes d'agrégation qui suivent. Elles ont un meilleur pouvoir prédictif. Plusieurs arbres sont construits à l'aide de données ; ils sont ensuite agrégés, ce qui limite notamment le risque de sur-apprentissage.

4.4.3 Les forêts aléatoires de survie

Les arbres de décisions sont très instables et ont de faibles capacités de prédictions. Ainsi nous utilisons les algorithmes d'agrégation, comme les forêts aléatoires qui ont une capacité de prédiction plus efficace sur les nouvelles données.

La courbe ci-dessous présente le taux d'erreur OOB en fonction du nombre d'arbres de survie utilisés pour le modèle.

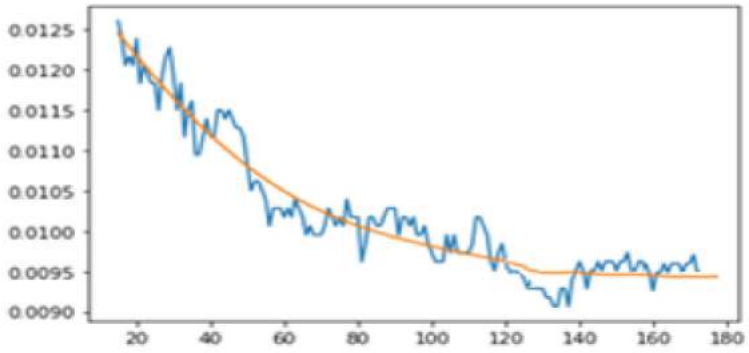


FIGURE 4.5 – Taux d'erreur OOB

Nous cherchons le moment où celui-ci est minimal et se stabilise. C'est au niveau de 130 arbres que le taux se stabilise. Par sécurité, nous en ajouterons 20 et utiliseront donc 150 arbres pour notre modèle.

Les forêts aléatoires ont permis d'améliorer considérablement les résultats obtenus avec les arbres.
A partir de la matrice de confusion du modèle, nous calculons la précision, le rappel et le f1-score qui fournissent d'excellents résultats.

```
[[2975   32]
 [   0  869]]
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      3007
           1       0.96      1.00      0.98       869

    accuracy                           0.99      3876
   macro avg       0.98      0.99      0.99      3876
weighted avg       0.99      0.99      0.99      3876
```

FIGURE 4.6 – Résultats obtenus avec les forêts aléatoires

Les calculs sont effectués pour chacune des catégories.
La classe 0 correspond aux emprunteurs qui n'ont pas subi le sinistre Perte d'emploi pendant la durée d'observation et la classe 1 correspond à des emprunteurs qui ont perdu leur emploi. Le support correspond au nombre d'observations dans chacune des catégories.

Résultats obtenus Nous notons d'excellent résultats avec cette méthode, tant au niveau de la précision, du rappel et du f1-score, et ce pour les deux catégories. En effet, les valeurs des indicateurs sont entre 0,96 et 1 et donc, le modèle, lors de la vérification sur la base d'entraînement, a obtenu des résultats très proches de la réalité. La moyenne des indicateurs est de 0,99.

4.4.4 Le Gradient Boosting

Au vu de la qualité des modèles d'agrégation, nous choisissons d'en implémenter un second type, le Gradient Tree Boosting. C'est un algorithme beaucoup plus complexe que le précédent et très coûteux en temps de calcul. Il s'avère être chronophage lors de sa mise en place.

A partir de la matrice de confusion du modèle, nous calculons la précision, le rappel et le f1-score qui fournissent de nouveau d'excellents résultats.

```
[[2961   46]
 [   0  869]]
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      3007
           1       0.95      1.00      0.97       869

    accuracy                           0.99      3876
   macro avg       0.97      0.99      0.98      3876
weighted avg       0.99      0.99      0.99      3876
```

FIGURE 4.7 – Résultats obtenus avec le Gradient Boosting

Les supports restent naturellement inchangés. Les valeurs des indicateurs sont bonnes, de moyenne 0,98.
Les forêts aléatoires et le Gradient Boosting fournissent tous deux d'excellents résultats, bien que légèrement meilleurs pour les forêts aléatoires.

4.4.5 L'importance des variables

Les modèles agrégés ont une meilleure précision et capacité de prédiction. Ainsi ils expliquent le lien entre les variables explicatives et la variable cible. Bien que ces modèles soient difficilement interprétables, il est possible d'extraire des modèles agrégés pour l'influence de chaque variable. Cela sera extrêmement utile, notamment afin d'émettre des recommandations et des points de vigilance sur le produit.

Une fois encore l'indice Gini est utilisé comme critère d'hétérogénéité, pour déterminer l'importance des variables. L'importance d'une variable correspond à la somme des décroissances d'hétérogénéité induites lorsqu'une variable intervient pour définir une segmentation d'un nœud. Plus une variable est utilisée à de nombreuse reprise pour segmenter les nœuds et ainsi réduire leur hétérogénéité, plus elle sera importante.

En comparant l'importance des variables déduites des Forêts aléatoires et du modèle XG Boost, les résultats sont très proches.

Résultats obtenus : Nous constatons que sur l'ensemble du portefeuille :

— la variable la plus importante est clairement la catégorie socioprofessionnelle. Les employés du secteur privé sont nettement plus à risque que les employés du secteur public ou parapublic
— Le nom du partenaire est la seconde variable la plus influente
— Les prêts financés par le Partenaire1 semblent bien plus risqués que les autres partenaires. Néanmoins, étant donné le très faible historique dont nous disposons sur les autres partenaires, nous ne considérons pas ce résultat comme pertinent
— Le montant assuré intervient ensuite négativement, plus il est élevé moins le risque ne l'est.
— Le sexe a aussi une influence non négligeable, les hommes sont une population plus à risque que la population féminine. Nous rappelons qu'une distinction tarifaire sur le sexe n'est pas autorisée, mais il est tout de même très intéressant de le remarquer, afin de rester vigilant quant à la proportion d'hommes dans le portefeuille
— La variable type de prêt nous indique que les prêts CQS ont une influence positive contrairement aux prêts secondaires, les *Delega*, qui augmentent le risque.

— La durée du contrat et l'âge arrivent en dernière position. Ils influent négativement, plus ils sont élevés, moins l'assuré est à risque de Perte d'emploi.

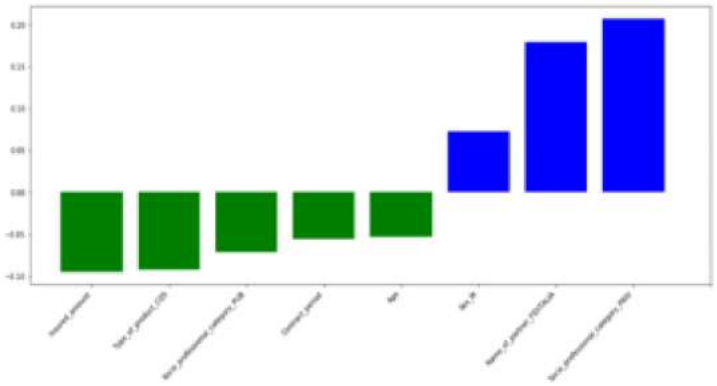


FIGURE 4.8 – Importance des variables sur l'ensemble du Portfeuille

La catégorie socio-professionnelle est indubitablement la variable la plus discriminante. De ce fait, afin d'affiner nos résultats nous effectuons une étude semblable en divisant le portefeuille en trois catégories selon le secteur d'activité. Alors, nous déterminons, avec les forêts aléatoires et le Gradient Boosting, l'importances des autres variables par rapport à la variable cible.

Nous obtenons des résultats très différents selon les catégories socio-professionelles.

Nous notons que pour la population des employés du secteur privé, le risque diminue avec l'âge, alors que pour la population des employés du secteur public il augmente.
Pour le secteur privé, les contrats les plus risqués sont ceux dont le montant assuré est faible, alors que, pour le secteur public, le montant assuré n'a aucune importance.

Globalement, quel que soit la catégorie, nous avons les conclusions suivantes :
- La population masculine est une population plus à risque que la population féminine
- Les prêts Delega sont des types plus à risque que les prêts CQS

Ci-dessous les schémas illustrant, les résultats de notre étude.

Secteur privé

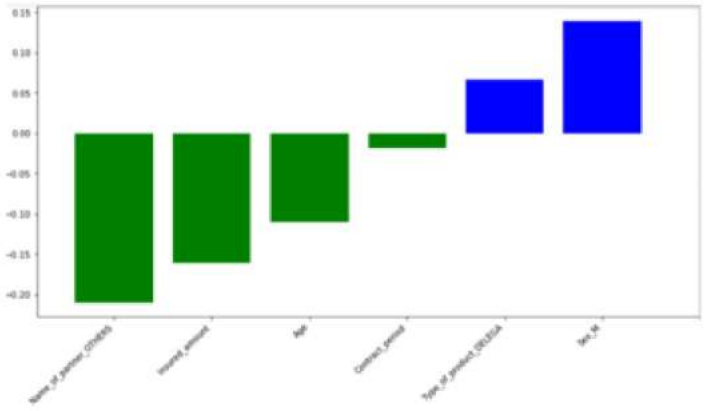


FIGURE 4.9 – Importance des variables - Secteur privé

Secteur public

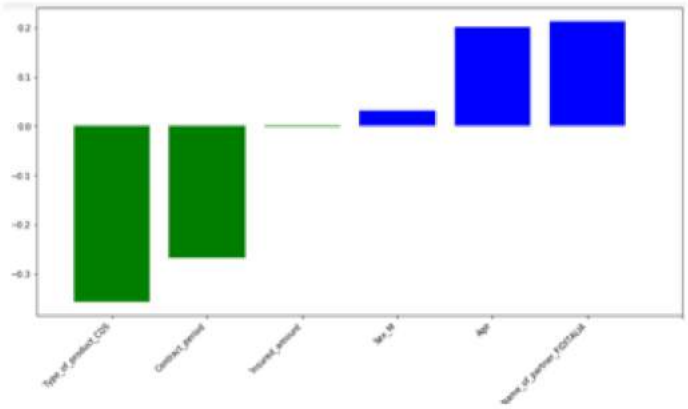


FIGURE 4.10 – Importance des variables - Secteur public

Secteur para-public

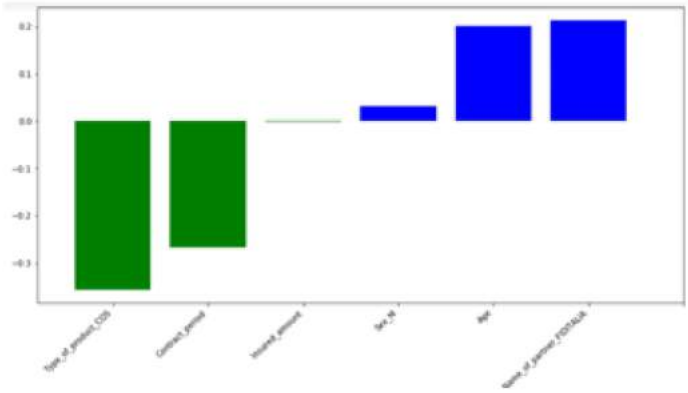


FIGURE 4.11 – Importance des variables - Secteur para-public

Cette partie nous a permis d'explorer les algorithmes de Machine Learning adaptés aux données censurées dont nous disposons. Plusieurs critères, tels que la précision, le rappel ou le f1-score, nous permettent de déterminer si les modèles utilisés sont ou non performants.Nous retenons les arbres de survie, les forêts aléatoires et le Gradient Boosting.
Notre base souffre d'un problème de déséquilibre des données, nous l'avons donc rééchantillonée, sans quoi, nous ne pouvons appliquer les méthodes ci-dessus.
Nous extrayons des modèles l'importance des varibles. Il est interessant de noter que nous obtenons des résultats très en ligne avec ceux obtenus à la partie précédente.

Nous retenons l'algorithme des forêts aléatoires, qui nous fournit d'excellents résultats, et nous l'utilisons pour prédire la perte d'emploi ou non d'un assuré du portefeuille au vu de ses caractéristiques. Une fois encore, la catégorie socio-professionelle est la variable la plus discriminante. Nous nous interessons donc aux variables qui suivent, après avoir partagé notre portefeuille selon les catégories

socio-professionelles. Les tendances sont différentes concernant l'âge et le montant emprunté. Mais une fois encore nous concluons que le hommes et les emprunteurs d'un Delega sont plus risqués.

Conclusion

En assurance, éviter les erreurs de financement constitue un objectif primordial pour assurer la pérénité des produits, la modélisation des risques est l'outil principal pour y parvenir. Des études actuarielles sont menées régulièrement afin de surveiller l'évolution des produits sur le marché.

L'objectif principal de ce mémoire est de modéliser le taux d'incidence de la perte d'emploi sur un produit spécifique, le *Cessione del Quinto*, en appliquant différentes méthodes de modélisation et d'ajustement. Nous voulons mieux analyser et comprendre les spécificités autour de ce produit d'assurance emprunteur italien.

L'analyse des statistiques a révélé des disparités dans la répartition des individus. Dans un premier temps, les définitions classiques de la prévalence et du taux d'incidence nous ont permis de calculer les taux d'incidence de l'évènement d'intérêt par année. Pour une analyse plus précise, nous avons segmenté la population en trois catégories socio-professionelle (privé, public et para-public). Nos données sur l'année 2019 sont incomplètes, nous avons donc jugé nécessaire d'adapter ces résultats, en incluant les IBNRs. La méthode d'estimation des réserves de Chain Ladder, basée sur des triangles de développement, nous a permis d'estimer les IBNRs. Nous les avons alors inclus dans nos données et avons calculé les taux. Cette première étape nous a permis d'identifier la catégorie la plus risquée : les employés du secteur privé sont 10 fois plus sujets à la perte d'emploi.

Les chapitres suivants présentent différentes méthodes : les méthodes d'analyse de survie et le modèle de Cox et celles de *Machine Learning*.

Dans un premier temps, les calculs de taux d'incidence selon la méthode de Kaplan Meier et le modèle de Cox nous ont apporté de nombreuses indications concernant le risque perte d'emploi en fonction de la catégorie socio-professionnelle, du sexe et l'âge. Ces méthodes sont communément utilisées en analyse de survie, car elle permettent de prendre en considération la censure, qui survient lorsque nous cessons de suivre la population, avant la survenance de l'évènement d'intérêt. En raison de certaines irrégularités, l'application d'un lissage sur les taux bruts de perte d'emploi obtenus s'est montrée nécessaires. Un lissage bien adapté permet à la fois d'éviter d'accorder trop d'importance à des données présentant un caractère peu significatif tout en tenant rigueur de la fidélité aux données brutes. Nous avons utilisé la méthode de Whittaker-Henderson qui tient compte de ces deux paramètres et prend également en considération les poids des données brutes. Nous déduisons des taux d'incidence lissés en fonction des différentes modalités.

L'étape finale de cette étude consiste à évaluer l'impact des différentes covariables, qui sont les caractéristiques des assurés, sur la variable cible, la survenue de la perte d'emploi, avec des méthodes de *Machine Learning*. Des récentes études ont permis d'étendre le *Machine Learning* à l'analyse de survie. Nous utilisons des algorithmes supervisés afin de modéliser l'incidence de l'évènement et effectuer des

prédictions. Afin d'obtenir les meilleures prédictions possibles, plusieurs algorithmes ont été testés, via un processus très rigoureux et méthodique. Nous avons finalement retenu les résultats des méthodes les plus performantes dans le cadre de notre étude : les forêts aléatoires et le *Gradient Boosting*.

À l'issue de cette étude, il est établi que l'apprentissage automatique est en mesure de fournir un complément pertinent. Les résultats sont axés dans les mêmes directions que ceux des méthodes précédentes. Nous distinguons les covariables les plus impactantes. Nous identifions la catégorie socio-professionelle, le sexe et le produit. Une approche de Machine Learning est un complément utile aux approches classiques, pour effectuer une analyse plus poussée et pertinente. En effet, elle permet d'extraire, pour chaque individu en portefeuille, la courbe de survie et ainsi d'identifier chaque individu à risque.

Cette analyse est enrichissante et permet diverses applications actuarielles, telle que l'étude du provisionnement nécessaire par produit d'assurance. Ainsi, les résultats obtenus permettent de vérifier si les méthodes de provisionnement actuelles sont pertinentes et si elles exigent un niveau de provisions suffisant. N'oublions pas que dans le cadre de la réforme réglementaire européenne du monde des assurances, Solvabilité II, il est primordial de s'assurer que les réserves actuelles permettent de couvrir l'ensemble des engagements envers les assurés du portefeuille.

Bibliographie

1. **JAKOBOWIXZ E.** [2018] Python pour le data scientist : Des bases du langage au machine learning
2. **THERNEAU T., GRAMBASCH M.** [2013] Modeling survival data
3. **NIKULIN M., WU H.** [2016] The Cox Model and Its Applications
4. **MILLER R.G.** [1998] Survival Analysis
5. **O'QUIGLEY J.** [2008] Proportional Hazards Regression
6. **AALEN O., GJESSING H.** [2008] Survival and Event History Analysis : A Process Point of View
7. **GERON A.** [2017] Machine Learning avec Scikit-Learn
8. **BATTY M.** [2019] Big data et machine learning
9. **UPINGO J.** [2016] Python for Probability, Statistics, and Machine Learning
10. **SHARMA N.** [2018] XGBoost - The Extreme Gradient Boosting for Mining Applications
11. **PLANCHET F.** [2020] Modèles de durée : Méthodes de lissage et d'ajustement
12. **PLANCHET F.** [2001] Critères de validation : Aspects méthodologiques
13. **PLANCHET F.** [2001] Modélisation statistique des phénomènes de durée
14. **PARTRAT C.** [2007] Provisionnement technique en Assurance non-vie
15. **GAO G.** [2018] Bayesian Claims Reserving Methods in Non-life Insurance with Stan : An Introduction
16. **BUSSY S.** [2019] Introduction of high-dimensional interpretable machine learning models and their applications
17. **MACHIN D.** [1995] Survival Analysis - A practical Approach
18. **OTTENWAELTER P.** [2014] Etude des lois d'incidence des sinistres décès et perte d'emploi des contrats Cessione del Quinto
19. **COX D.R.,OAKES D.** [2018] Analysis of survival data
20. **ALISON D. P.** [1984] Event history and survival analysis
21. **ADELE A.** [2016] Construction d'une table d'expérience pour le maintien en incapacité
22. **KOYE G.** [2020] Comparaison des méthodes classiques et alternatives avec le machine learning pour la construction d'une table de mortalité d'expérience Best Estimate
23. **JACOBOWICZ E.** [2019] Introduction au machine learning
24. **COURVILLE A.** [2015] Deep learning
25. **HERBRETEAU A.** [2016] Construction d'un outil de suivi de la rentabilité pour le portefeuille italien de la Cessione del Quinto

26. **BIZOUARD Y.** [2015] Modélisation et provisionnement du risque de Perte d'Emploi dans le cadre de l'Assurance Emprunteur
27. **CARPENTIER F.** [2018] Introduction à l'analyse de survie
28. **POGGI J.M.** [2019] Les forêts aléatoires avec R
29. **SHARMA N.** [2018] XGBoost. The Extreme Gradient Boosting for Mining Applications

Table des figures

Liste des tableaux

Liste des acronymes

CQS Cessione del Quinto dello Stipendio

CQP Cessione del Quinto della Pensione

CQ Cessione del Quinto

SGI Societe generale Insurance

INPDAP Instituto Nazionale di Previdenza per i dipendenti dell'Amministrazione Pubblica

TFR Trattamento di Fine Rapporto

TAN taux annuel nominal

LoE Loss of employment

IBNR Incured but not reported

ML Machine Learning

MAPE Mean Absolute Percentage Error

CART Classification And Regression Trees

OOB Out of bag

TP True positive

FP False positive

FN False negative

TN True negative

XGB eXtreme Gradient Boosting