duas
Diplôme Universitaire des Actuaires de Strasbourg

UFR Mathématique Informatique Strasbourg

Mémoire présenté devant

l'UFR de Mathématique et Informatique

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

et l'admission à l'Institut des Actuaires

le 01 / 12 / 2021

Par : Samuel Boisadam

Titre: Apport des méthodes de *Machine Learning* sur la modélisation des taux de Résiliation en Assurance Emprunteur

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

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

Signature :

ADDACTIS France
Espace WeWork
4, rue Jules Lefebvre - 75009 Paris
Tél. 01 56 89 07 70
Siren 413 611 344 - NAF 7022 Z
TVA Intracommunautaire FR53 413 611 344
SAS au capital de 320 000 €

Membres du jury de l'Unistra :

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

Entreprise : ADDACTIS France

Directeur de mémoire en entreprise

Nom : Antoine BRUN

Invité :

Nom :

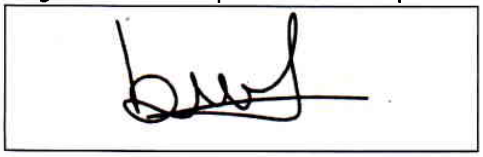
Signature :

Jury de l'Institut des Actuaires :

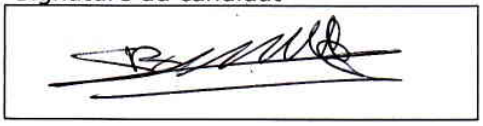
M. KELLE VIGON
F. HENGE

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

Signature du responsable entreprise



Signature du candidat



Secrétariat : Mme Stéphanie Richard

Bibliothèque : Mme Christine Disdier

Remerciements

Tout d'abord, je tiens à remercier Antoine BRUN, mon tuteur, pour son accompagnement et sa grande implication tout au long de ce mémoire. Ces remerciements s'appliquent également à Fabien VETILLARD, Nabil RACHDI et Stéphanie LAMI pour toutes les pistes de réflexion et relectures qu'ils ont pu apporter.

Je suis particulièrement reconnaissant à Khedija ABDELMOULA CLAVERIE pour son accueil au sein de l'équipe *Actuarial Consulting*. Je remercie également tous les membres de l'entité Addactis pour avoir contribué au bon déroulement de ce travail.

J'aimerai ensuite remercier l'ensemble du corps professoral du DUAS pour la qualité de l'enseignement dispensé ainsi que Jean MODRY, mon tuteur universitaire, pour son encadrement, ses relectures et ses conseils.

Je souhaite également remercier mes collègues alternants et plus particulièrement Bérengère DAYNAC, Marie LEMARIÉ et Floriane BUFFET pour leur bonne humeur quotidienne.

Enfin, j'ai une pensée particulière pour mes proches et notamment mes parents et ma soeur pour leur soutien constant tout au long de mes études.

Résumé

Mots clés : assurance emprunteur, résiliation, modélisation, Kaplan-Meier, Whittaker-Henderson, arbre de décision, forêts aléatoires, XGBoost, forêts aléatoires de survie.

L'assurance emprunteur intervient lors de la contraction d'un prêt. Suite aux récentes évolutions législatives, et notamment l'amendement Bourquin, une résiliation annuelle sans aucune modification du prêt sous-jacent est désormais possible. Cette forme de résiliation vient s'ajouter à celle qui entraîne le rachat du crédit.

Avec l'essor des méthodes de *Machine Learning*, l'objectif de ce mémoire est de confronter ces méthodes à une méthode actuarielle classique, tant d'un point de vue technique que financier, en se fondant sur une base de données possédant une quantité d'informations limitée.

Dans une première partie, nous détaillerons les évolutions législatives et caractéristiques des contrats emprunteur. Ces caractéristiques intègrent les spécificités du marché mais aussi les garanties ainsi que les différents modes de remboursement et de tarification. Dans une seconde partie, les taux de résiliation seront modélisés à l'aide de l'estimateur de Kaplan-Meier et du lissage de Whittaker-Henderson. Ces taux serviront de base de comparaison pour les taux modélisés à l'aide des méthodes d'arbre de décision, de *Random Forest* et XGBoost. Dans une dernière partie, une étude d'impacts sera réalisée au travers de l'analyse de la valeur actuelle probable des résultats futurs. Celle-ci permettra d'évaluer la pertinence des méthodes de *Machine Learning*. Enfin, un intérêt particulier sera apporté à l'influence des variables, à la notion de censure pour les méthodes de *Machine Learning* avec l'implémentation de la méthode *Random Survival Forest* ainsi qu'aux opportunités offertes en matière de segmentation des tarifs.

Abstract

Key words : credit insurance, termination, modeling, Kaplan-Meier, Whittaker-Henderson, decision tree, Random Forest, XGBoost.

The credit insurance is involved when taking out a loan. Following the recent legislative changes, and in particular the Bourquin amendment, an annual termination without any modification of the underlying loan is now possible. This form of termination is in addition to the one that leads to credit buyout.

With the development of Machine Learning methods, this work aims to confront these methods with a classical actuarial method, both from a technical and financial point of view, based on a database with a limited amount of information.

Firstly, the legislative evolutions and characteristics of the loan contracts will be detailed. These characteristics integrate the market specificities but also guarantees, different modes of reimbursement and pricing. Secondly, the termination rates will be modeled using the Kaplan-Meier estimator and Whittaker-Henderson smoothing. These rates will be used as a basis for comparison with rates modeled using the decision tree, Random Forest and XGBoost methods. Lastly, an impact study will be carried out through the analysis of the expected present value of future results. This will allow to evaluate the revelance of Machine Learning methods. Finally, special interest will be paid to the influence of variables, to the notion of censorship for Machine Learning methods with the implementation of the Random Survival Forest method, and to the opportunities offered in terms of prices segmentation.

Note de synthèse

Mots clés : assurance emprunteur, résiliation, modélisation, Kaplan-Meier, Whittaker-Henderson, arbre de décision, forêts aléatoires, XGBoost.

Le marché de l'assurance emprunteur en France fait l'objet de réformes visant à augmenter la concurrence et réduire les marges pratiquées. Les lois Lagarde (2010) et Hamon (2014) puis l'amendement Bourquin (2017) permettent désormais une résiliation à tout moment la première année puis à chaque date d'anniversaire du contrat. Ces réformes entraînant l'apparition d'une résiliation du contrat emprunteur sans modification du prêt auquel il est rattaché, une révision des lois de rachat, intégrant cette nouvelle composante, permettrait d'appréhender au mieux ce risque.

En parallèle, l'essor des méthodes de *Machine Learning* offre de nouvelles opportunités de modélisation. Toutefois, ces méthodes, que l'on retrouve principalement dans le milieu de l'assurance non-vie, ne sont pas communes dans le secteur de l'assurance emprunteur et leur pertinence reste à démontrer. En effet, il est nécessaire que les résultats obtenus justifient d'un intérêt suffisant pour pallier les contraintes supplémentaires, techniques et temporelles, apportées par ces méthodes.

Dans le cadre d'une base de données aux informations limitées, l'objectif de cette étude sera donc de proposer une modélisation du risque de résiliation selon une méthode actuarielle classique puis selon différentes méthodes de *Machine Learning*. Après étude de la robustesse de ces dernières, une comparaison avec la méthode actuarielle classique sera réalisée à l'aide d'un indicateur basé sur la valeur actuelle probable des contrats. L'ajout de la notion de censure dans les méthodes de *Machine Learning* sera également étudiée. Enfin, une attention particulière sera accordée aux informations complémentaires apportées par ces dernières et notamment l'importance des variables.

Concernant la modélisation de la loi de résiliation par une méthode actuarielle classique, le choix s'est porté sur l'utilisation de l'estimateur des taux de survie de Kaplan-Meier. Un prêt étant généralement souscrit sur une durée pouvant atteindre 25 ou 30 ans, cet estimateur est celui qui permet de prendre le plus en compte le phénomène de censure lié à la période

d'observation restreinte. La base de données utilisée pour cette étude ne permettant pas une étude statistique rigoureuse, le choix a été fait de procéder à une méthode *bootstrap* afin d'obtenir les intervalles de confiance autour des estimations des taux de survie. Les bornes des intervalles de confiance sont obtenues après application de la formule de l'estimateur de Kaplan-Meier à un grand nombre d'échantillons de données. Ces échantillons sont obtenus à l'aide d'un tirage aléatoire avec remise sur l'échantillon de départ. Ces intervalles sont ensuite utilisés afin de valider ou invalider les lissages de Whittaker-Henderson réalisés. Combinaison linéaire entre un critère de régularité et un critère de fidélité, cette méthode de lissage vise à effacer le plus possible les spécificités du jeu de données tout en y restant fidèle. Une fois les paramètres du lissage retenus sur un critère de prudence ainsi que le passage des taux de survie lissés aux taux de résiliation effectué, la loi de résiliation est entièrement obtenue en appliquant la fonction de survie de Weibull au niveau de la dernière ancienneté disponible. Le choix de cette fonction pour prolonger les taux résulte de l'hypothèse selon laquelle les taux de résiliations sont décroissants et tendent vers 0 % passée la 8ème année d'ancienneté du prêt. En complément de la loi obtenue, une étude selon une segmentation en trois classes d'âge a été réalisée. Le choix de cette segmentation a pour objectif d'évaluer l'influence de l'âge à la souscription sur la résiliation. Ainsi, les lois obtenues sont les suivantes :

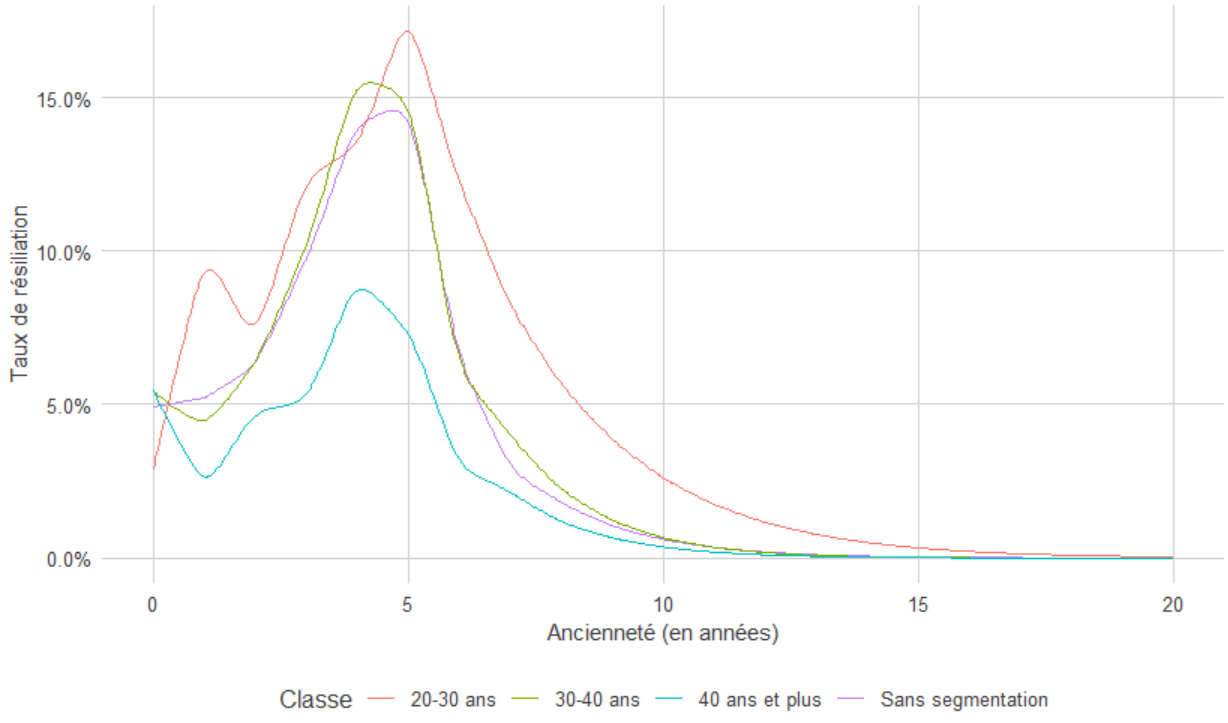


FIGURE 1 – Taux de résiliation obtenus par la méthode actuarielle

Des différences de comportement peuvent être déduites des lois obtenues. Plus l'âge à la souscription augmente, moins les assurés sont enclins à résilier leur contrat emprunteur. Ce phénomène pourrait s'expliquer par deux facteurs. D'une part, plus les profils sont âgés plus ils sont exposés au risque de décès et d'arrêt de travail et donc plus un changement de contrat serait coûteux. D'autre part, les profils les plus jeunes manifesteraient un intérêt plus important à effectuer des démarches pour résilier leur contrat.

Concernant la modélisation de la loi de résiliation par des méthodes de *Machine Learning*, les modèles suivants ont été étudiés :

- modèle d'arbre de décision ;
- modèle de forêts aléatoires (*Random Forest*) ;
- modèle *eXtreme Gradient Boosting* (XGBoost).

Le modèle d'arbre de décision est fondé sur la méthode CART (*Classification And Regression Tree*). Cette méthode se base sur la classification des données par une suite de tests réalisés sur les variables explicatives. À terme, la méthode présente les données classifiées sous forme d'arbre. Le modèle de *Random Forest* repose quant à lui sur l'apprentissage et l'agrégation d'un ensemble d'arbres de décision. Enfin, le modèle XGBoost est fondé sur l'amélioration d'un arbre de décision par itérations successives.

Le même processus d'optimisation a été adopté pour les trois méthodes. Après une séparation de la base de données en une base d'entraînement et une base de test, les hyper-paramètres optimaux de chaque modèle ont été obtenus à l'aide de la méthode *Grid Search*, implémentée sous Python. Cette méthode consiste en une recherche exhaustive de la meilleure combinaison possible d'hyper-paramètres parmi toutes celles testées. Compte tenu de la faible taille de la base de données et afin de s'abstenir de la création d'une base de validation, la méthode *Grid Search* fait appel au concept de *cross validation*. Ce principe consiste à séparer la base d'entraînement en plusieurs blocs et d'utiliser chaque bloc comme base de validation après entraînement du modèle sur les autres blocs.

Une fois les modèles ajustés sur la base d'entraînement à l'aide des hyper-paramètres optimaux, il est possible de calculer des métriques en appliquant le modèle sur la base de test. Ces métriques permettent d'évaluer la précision globale du modèle mais également de déterminer le type d'erreurs réalisées. Les métriques agrégées que sont l'AUC ROC et l'AUC P-R sont celles qui, parmi toutes les métriques calculées, permettent de mieux mettre en avant le lien entre gain en performance et gain en précision. Pour les trois modèles étudiés, les résultats obtenus sont les suivants :

Modèle	AUC ROC	Gain de performance	AUC P-R	Gain de précision
Decision Tree	89,83 %		82,64 %	
Random Forest	90,67 %	+ 0,94 %	86,45 %	+ 4,61 %
XGBoost	93,45 %	+ 4,03 %	89,28 %	+ 8,03 %

TABLE 1 – Comparaison des métriques agrégées

Il ressort de ce tableau que le modèle XGBoost est en tout point plus performant que les deux autres modèles. Toutefois, afin de s'assurer que ces résultats ne sont pas dus au hasard, une étude complémentaire a été réalisée. Le processus d'optimisation a été appliqué à une centaine de séparations différentes de la base de données. Pour chaque séparation, toutes les métriques sont calculées et stockées. Les répartitions obtenues sont alors les suivantes :

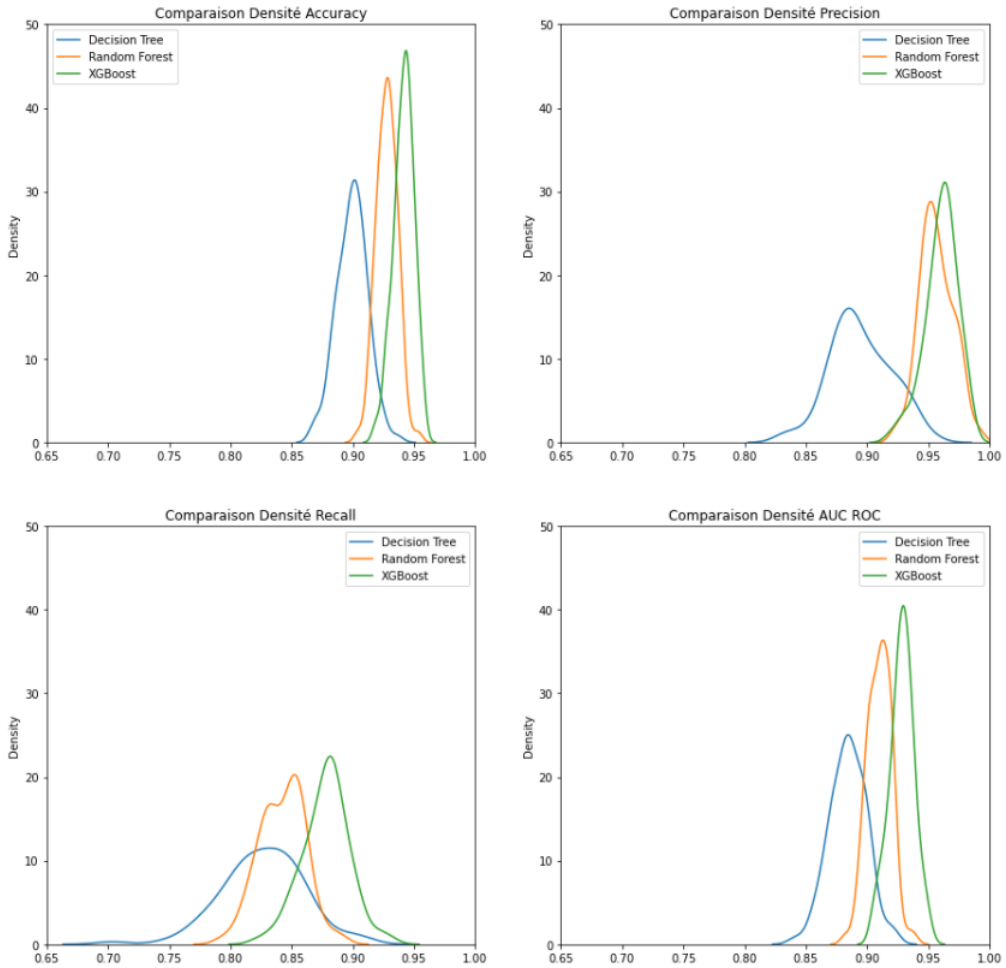


FIGURE 2 – Densités des différentes métriques de performance

Les résultats ci-dessus confirment ceux obtenus à l'aide des métriques agrégées. Le modèle XGBoost est celui ayant les meilleures performances. Ces densités transmettent également des informations concernant la sensibilité des modèles à l'aléatoire de la séparation en base

d'entraînement et base de test. Un faible écart-type est représentatif d'un modèle peu soumis à cette composante aléatoire ce qui est le cas pour le modèle XGBoost. En considérant l'ensemble des résultats observés, si un choix devait être fait parmi ces trois méthodes selon un critère de performance pure, la décision devrait s'orienter vers la méthode XGBoost. Toutefois, si un arbitrage doit être réalisé entre performance et temps de calcul, la méthode par arbre de décision est la plus pertinente. Cette méthode présente des performances proches de celles des méthodes les plus efficientes tout en offrant un temps de calcul rapide.

Une fois les taux de résiliation obtenus à l'aide des méthodes de *Machine Learning*, il est possible de les comparer avec la méthode actuarielle classique :

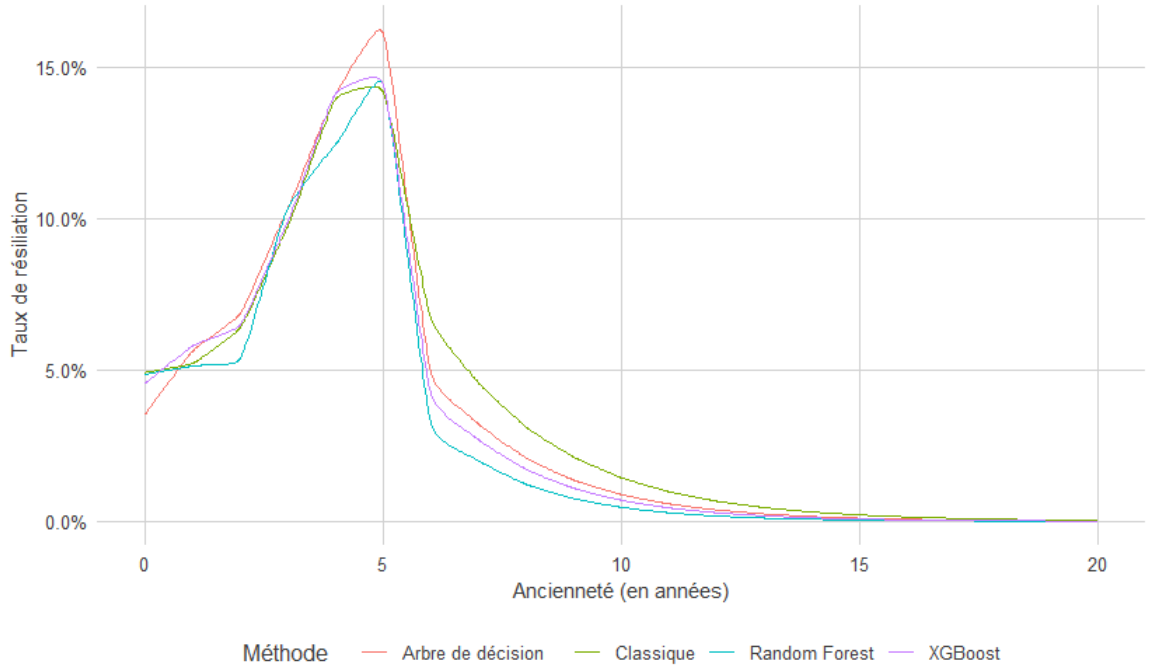


FIGURE 3 – Taux de résiliation pour les différentes méthodes considérées

Bien que les taux de résiliation présentent une allure similaire, il n'est pas possible de déterminer graphiquement si l'une des méthodes se détache des autres taux d'un point de vue économique. C'est pour cette raison qu'un indicateur de performance est calculé. Cet indicateur, basé sur la valeur actuelle probable des résultats nets d'un portefeuille en situation de *run-off*, est le ratio I déterminé de la manière suivante :

$$I = \frac{VAP_{méthode\ comparée}}{VAP_{méthode\ actuarielle}} - 1$$

Pour comparer les méthodes, deux scénarios ont été envisagés. Le premier est un portefeuille non segmenté où le profil-type est celui correspondant aux statistiques moyennes de la base de données complète. Le second scénario correspond à un portefeuille segmenté en trois classes d'âge, les mêmes que pour la modélisation actuarielle classique, où les profils-types sont ceux correspondants aux statistiques moyennes de chaque classe d'âge. Les résultats obtenus sont les suivants :

Méthode considérée	Scénario 1	Scénario 2
Actuarielle		
Arbre de décision	+ 0,46 %	+ 0,49 %
Random Forest	+ 2,37 %	+ 1,87 %
XGBoost	+ 0,70 %	+ 0,49 %

TABLE 2 – Comparaison des méthodes pour les scénarios 1 et 2

Les résultats montrent que, dans le cadre de cette étude, la méthode actuarielle classique est la méthode dont l'approche est la plus prudente. Afin d'essayer d'expliquer l'augmentation plus importante de la VAP pour le modèle *Random Forest*, un modèle de *Machine Learning* intégrant la censure est étudié. Basé sur le principe du *Random Forest* et après optimisation du modèle, les résultats obtenus sont les suivants :

Méthode considérée I	Scénario 1	Scénario 2
Random Forest	+ 2,37 %	+ 1,87 %
Random Survival Forest	+ 0,50 %	+ 0,13 %

TABLE 3 – Comparaison du modèle de *Random Survival Forest* pour les scénarios 1 et 2

De ces résultats, il est possible de déduire que pour cette étude, l'ajout de la notion de censure aux méthodes de *Machine Learning* permet d'accentuer l'approche prudente des taux de résiliation sans pour autant arriver au niveau de la méthode actuarielle classique.

Les méthodes de *Machine Learning* permettent également, en plus de modéliser les taux de résiliation des contrats d'assurance emprunteur, de déterminer l'importance des différentes variables dans l'explication du phénomène de résiliation. Ces influences sont un argument de taille pour envisager de nouvelles segmentations des taux de résiliation afin que ceux-ci soient les plus représentatifs du risque réel. Après étude de l'influence des variables pour chacune des méthodes, il s'avère que le montant emprunté est l'une de celle qui a la plus grande influence sur la détermination des taux de résiliation. Ainsi, une segmentation prenant en compte l'ancienneté du prêt et le montant emprunté permet d'aboutir aux taux suivants :

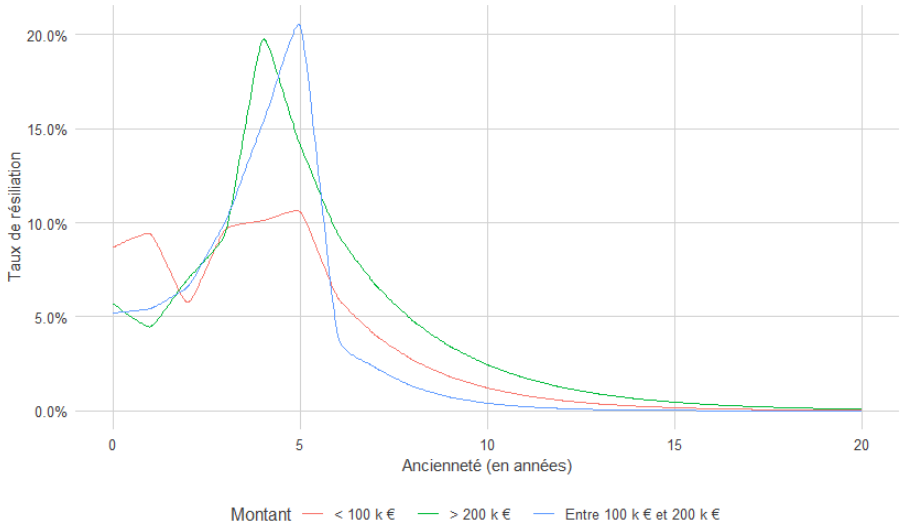


FIGURE 4 – Taux de résiliation segmentés par montant emprunté

De ces taux de résiliation se dégagent deux tendances. D'une part, les prêts dont le montant emprunté est inférieur à 100 000 €. De l'autre, les prêts dont le montant emprunté est supérieur à 100 000 €. Il est alors possible de faire des conjectures sur le comportement des assurés. Pour les prêts inférieurs à 100 000 €, la prime étant proportionnelle au montant emprunté, l'intérêt financier à résilier est moindre tandis que pour les prêts supérieurs à 100 000 €, la résiliation peut se justifier par une plus grande possibilité d'économie sur le montant de son assurance emprunteur.

En l'état, dans cette étude, il n'est pas pertinent de vouloir remplacer les méthodes actuarielles classiques par les méthodes de *Machine Learning* pour modéliser les taux de résiliation. Certes, ces dernières permettent d'appréhender précisément le risque de résiliation et apportent des fonctionnalités permettant une meilleure segmentation des taux. Toutefois, ces méthodes impliquent un coût temporel de mise en place non négligeable. De plus, ces méthodes sont également extrêmement sensibles aux données en entrée de modèle. Il faut donc accorder une attention particulière à la fois à la qualité mais aussi à la quantité des données. Enfin, même si cette étude ne permet pas d'aboutir à des résultats attestant d'un réel intérêt pour les modèles de *Machine Learning*, elle peut toutefois être à la base d'études complémentaires sur ces méthodes et sur l'intérêt de disposer d'un outil prédictif du risque de résiliation.

Note of synthesis

Key words : credit insurance, termination, modeling, Kaplan-Meier, Whittaker-Henderson, decision tree, Random Forest, XGBoost.

The loan insurance market in France is undergoing reforms aimed at increasing competition and reducing margins. The Lagarde law in 2010, the Hamon law in 2014 and the Bourquin amendment in 2017 now allow termination at any time during the first year and then on each anniversary date of the contract. As these reforms lead to the termination of the loan insurance contract without modification of the underlying credit, a revision of the buyout laws, integrating this new component, would allow this risk to be better understood.

In the same time, the development of Machine Learning methods offers new modeling possibilities. However, these methods, which are mainly found in the non-life insurance industry, are not common in the loan insurance sector and their relevance remains to be demonstrated. Indeed, it is necessary that the results obtained justify a sufficient interest to compensate for the additional technical and temporal constraints brought by these methods.

In the context of a database with limited information, this study aims to propose a modeling of the risk of termination according to a classical actuarial method and then according to different Machine Learning methods. After studying the robustness of the latter, a comparison with the classic acturial method will be carried out using an indicator based on the expected present value of the contracts. Lastly, special attention will be paid to the additional information provided by the Machine Learning methods and in particular the importance of the variables.

Concerning the modeling of the termination law by a classical actuarial method, the choice was made to use the Kaplan-Meier survival rate estimator. Since a loan is generally taken out over a period of up to 25 or 30 years, this estimator is the one most takes into account the censoring phenomenon associated with the limited observation period. As the database used for this study did not allow for a rigourous statistical study, the choice was made to use a boostrap method to obtain the confidence intervals around the survival rate estimates. The bounds of the confidence intervals are obtained after applying the Kaplan-Meier estimator

formula to a large number of data samples. These samples are obtained using a random draw with replacement from the starting sample. These intervals are then used to validate or invalidate the Whittaker-Hendeson smoothing performed. Combination between a regularity criterion and a fidelity criterion, this smoothing method aims at erasing as much as possible the specificities of the dataset while remaining faithful to it. Once the smoothing parameters have been selected on a conservative basis and the transformation for smoothed survival rates to termination rates has been made, the termination law is fully obtained by applying the Weibull survival function at the last available seniority level. The choice of this function to extend the rates results from the assumption that the termination rates are decreasing and tend towards 0 % after the 8th year of service. In addition to the law obtained, a study based on a segmentation into three age groups was carried out. This chosen to evaluate the influence of age at subscription on termination. Thus, the laws obtained are the following :

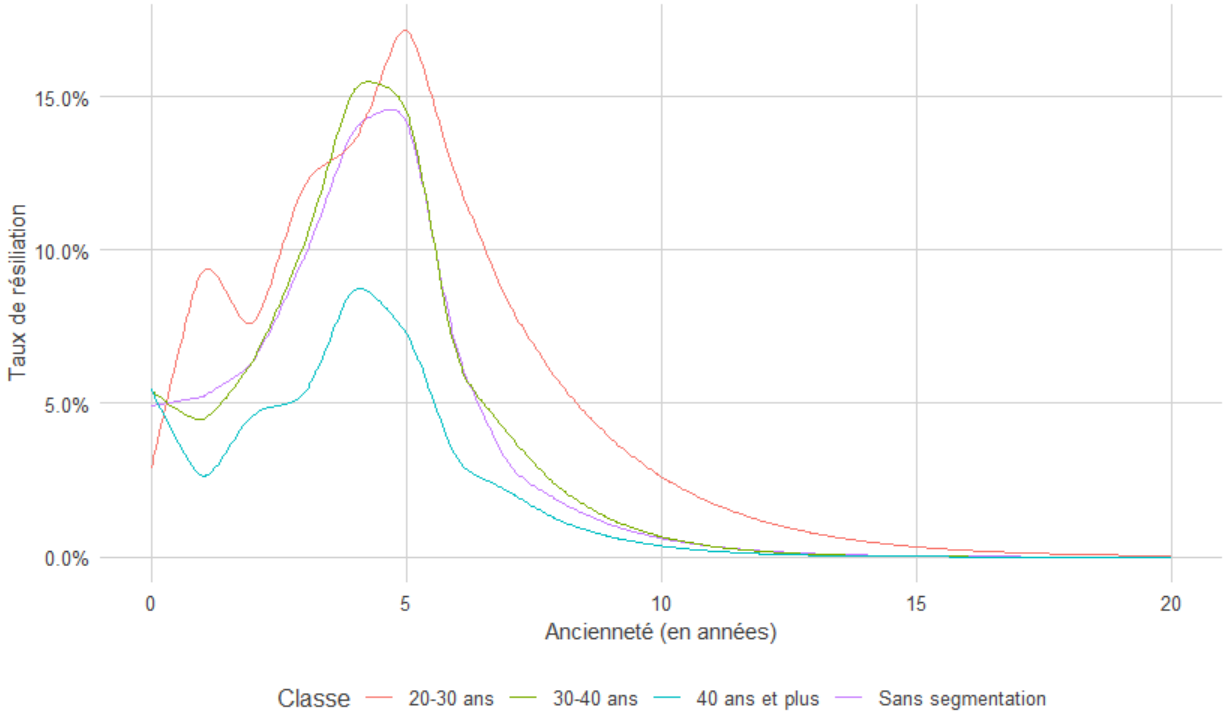


FIGURE 5 – Termination rates obtained with the actuarial method

Differences in behavior can be deduced from the laws obtained. The more the age at underwriting increases, the less policyholders are inclined to cancel their loan insurance contract. This phenomenon could be explained by two factors. On the one hand, the older the profiles, the more exposed they are to the risk of death and work stoppage and therefore the more expensive a change of contract would be. On the other hand, the youngest profiles would

show à greater interest in taking steps to terminate their contract.

Regarding the modeling of the termination law by Machine Learning methods, the following models have been studied :

- decision tree model ;
- Random Forest model ;
- eXtreme Gradient Boosting (XGBoost) model.

The decision tree model is based on the CART method (Classification And Regression Tree). This method relies on the classification of the data by a series of tests performed on the explanatory variables. Eventually, the method presents the classified data as a tree. The Random Forest model is based on the training and aggregation of a set of decision trees. Lastly, the XGBoost model is based on the improvement of decision tree by successive iterations.

The same optimization process was adopted for all three methods. After separating the database into a training and a test database, the optimal hyper-parameters of each model were obtained using the *Grid Search* method, implemented in Python. This method consists in an exhaustive search for the best possible combinaison of hyper-parameters among all those tested. Given the small size of the database and in order to avoid the creation of a validation database, the *Grid Search* method uses the concept of cross validation. This principle consists in separating the training base into several blocks and using each block as a validation base after training the model on the other blocks.

Once the models are fitted to the training base using the optimal hyper-parameters, metrics can be calculated by applying the model to the test base. These metrics allow us to evaluate the overall accuracy of the model but also to determine the type of errors made. The aggregate metrics of AUC ROC and AUC P-R are the ones that, among all the calculated metrics, best highlight the link between performance gain and accuracy gain. For the three models studied, the results obtained are as follows :

Model	AUC ROC	Performance gain	AUC P-R	Precision gain
Decision Tree	89,83 %		82,64 %	
Random Forest	90,67 %	+ 0,94 %	86,45 %	+ 4,61 %
XGBoost	93,45 %	+ 4,03 %	89,28 %	+ 8,03 %

TABLE 4 – Aggregate metrics comparison

The table reveals the XGBoost model performs better than the other two models in every aspect. However, in order to ensure that these results were not due to randomness, an

additional study was conducted. The optimization process has been applied to about 100 different separations of the database. For each seperation, all metrics are calculated and stored. The distribution obtained are then as follows :

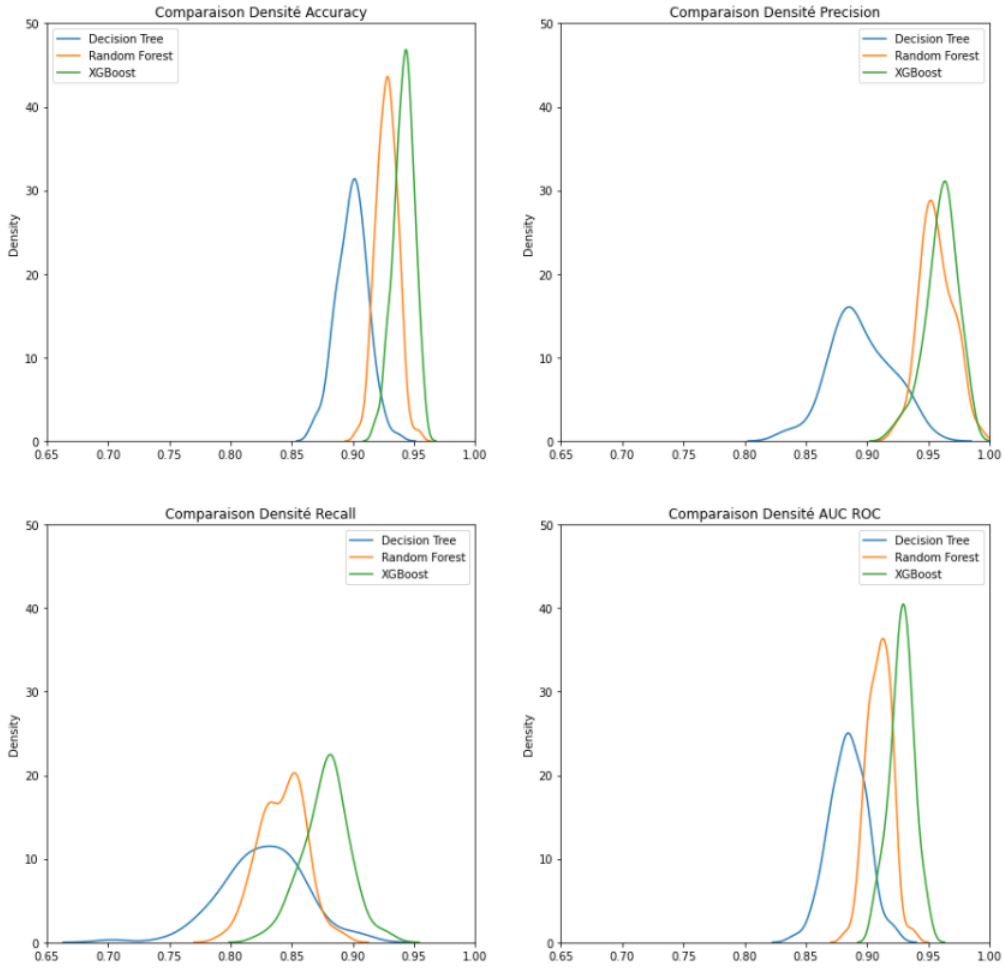


FIGURE 6 – Distribution of the different performance metrics

The above results confirm those obtained using the aggregate metrics. The XGBoost model is the one with the best performance. These densities also convey information about the sensitivity of the models to the randomness of the separation into training and test bases. A low standard deviation is reprentative of a model that is not very subject to this random component, which is the case for the XGBoost model. Considering all the observed results, if a choice had to be made among these three methods according to a pure performance criterion, the decision should be oriented towards the XGBoost method. However, if a trade-off has to be made between performance and computation time, the decision tree method is most revelant. This method presents performances close to those of the best performing methods while offering a quick computation time.

Once the termination rates are obtained using Machine Learning methods, it can be com-

pared with the classical method :

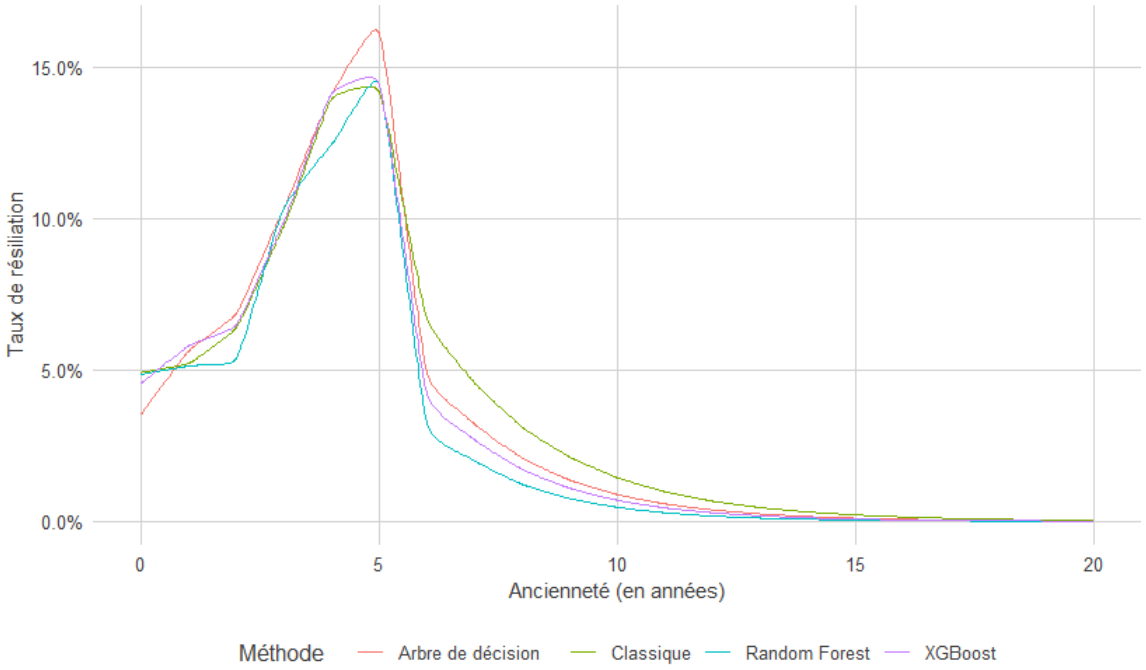


FIGURE 7 – Termination rates for the different methods considered

Although the termination rates look similar, it is not possible to determine graphically whether one method stands out from the others from an economic perspective. For this reason, a performance indicator is calculated. This indicator, based on the expected present value of the net results of portfolio in a run-off situation, is the ratio I determined as follows :

$$I = \frac{EPV_{compared\ method}}{EPV_{actuarial\ method}} - 1$$

To compare the methods, two scenarios are considered. The first is a non-segmented portfolio where the standard profile is the one corresponding to the average statistics of the complete database. The second scenario corresponds to a portfolio segmented into three age classes, the same as for the classical actuarial modeling, where the standard profiles are those corresponding to the average statistics of each age class. The results obtained are as follows :

Method	Scenario 1	Scenario 2
Actuarial		
Decision Tree	+ 0,46 %	+ 0,49 %
Random Forest	+ 2,37 %	+ 1,87 %
XGBoost	+ 0,70 %	+ 0,49 %

TABLE 5 – Methods comparison for scenarios 1 and 2

The results show that, in this study, the classical actuarial method is the most careful approach. In order to try to explain the higher increase in the EPV for the Random Forest model, a Machine Learning model incorporating censoring is studied. Based on the Random Forest principle and after optimisation of the model, the results obtained are as follows :

Ratio I	Scénario 1	Scénario 2
Random Forest	+ 2,37 %	+ 1,87 %
Random Survival Forest	+ 0,50 %	+ 0,13 %

TABLE 6 – Random Survival Forest method comparison for scenarios 1 and 2

From these results, it is possible to deduce that for this study, the addition of the notion of censoring to the Machine Learning methods allows to accentuate the careful approach of the termination rates without reaching the level of the classical actuarial method.

Machine Learning methods also permit, in addition to modeling the termination rates of loan insurance contracts, to determine the importance of the differents variables in the explanation of the termination phenomenon. These influences are a strong argument for considering new segmentations of termination rates so that they are most reprensentative of actual risk. After studying the influence of the variables for each of the methods, it turns out that the borrowed amount is one of the variables that has the greatest influence on determination of termination rates. Thus, a segmentation taking into account the age of the loan and the borrowed amount leads to the following rates :

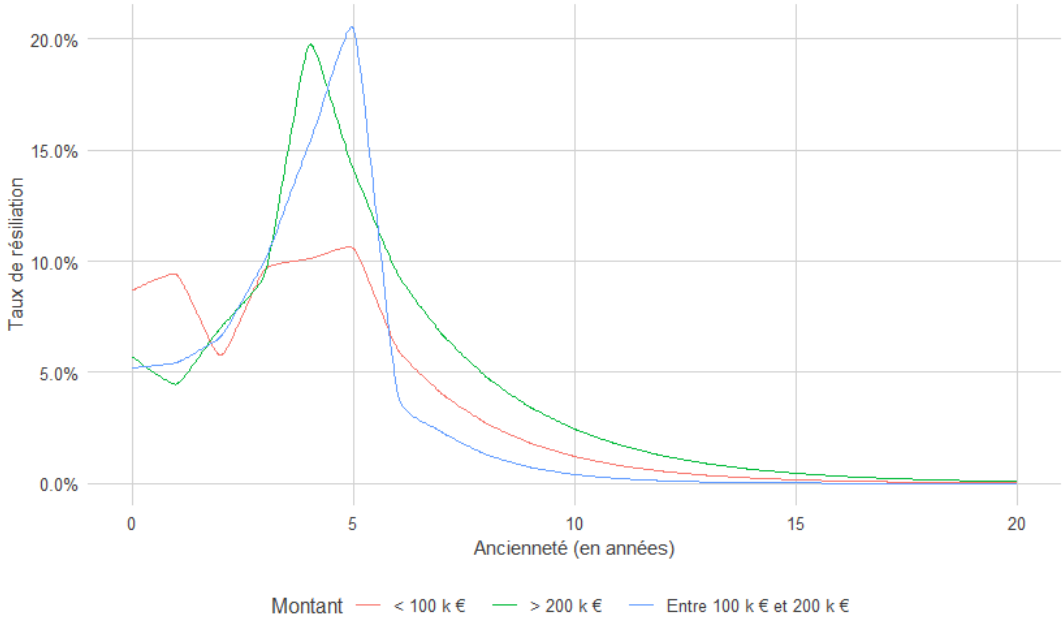


FIGURE 8 – Termination rates by borrowed amount

Two trends emerge from these termination rates. On the one hand, loans with a borrowed amount of less than 100,000 euros. On the other hand, loans with a borrowed amount of more than 100,000 euros. It is then possible to make conjectures about the behavior of the insured. For loans under 100,000 euros, the premium being proportionnal to the amount borrowed, the interest in terminate is less, whereas for loans over 100,000 euros, terminate may be justified by a greater possibility of saving on the amount of one's insurance.

As it stands, in this study, it is not revelant to replace traditional actuarial methods with Machine Learning methods to model termination rates. Admittedly, these methods allow for a precise understanding of the risk of termination and provide features that allow for a better segmentation of rates. However, these methods imply a significant time cost to implement. Moreover, these methods are also extremely sensitive to the model input data. Therefore, special attention must be paid to both the quality and the quantity of the data. To conclude, even if this study does not lead to results showing a real interest for Machine Learning methods, it can nevertheless be the basis for further studies on these methods and on the interest of having a predictive tool for the termination risk.

Introduction

La souscription d'une assurance emprunteur est généralement imposée par l'établissement de crédit lors de la contraction d'un prêt. Cette couverture revêt un double intérêt. D'une part, pour le débiteur, elle assure la conservation de ses biens en cas de perte de revenus et la non-transmission de dettes à ses héritiers en cas de décès. D'autre part, elle protège l'organisme bancaire contre le non remboursement du capital emprunté. Dans le cadre d'une étude menée par l'Inspection Générale des Finances en 2013 [IGF, 2013], l'assurance représenterait entre 5 % et 20 % du coût d'un crédit et serait un levier important d'économie pour les particuliers.

Selon la Fédération Française de l'Assurance [FFA, 2020], le marché de l'assurance emprunteur en France est majoritairement contrôlé par les bancassureurs qui ont perçu près de 89% des cotisations totales en 2019. Afin d'ouvrir ce marché à la concurrence, le gouvernement met en place depuis une dizaine d'années une série de lois facilitant le choix de son assurance emprunteur. En 2010, la loi Lagarde offre aux emprunteurs le libre choix de leur assurance lors de la souscription du crédit. Quatre ans plus tard, la promulgation de la loi Hamon permet la résiliation de son contrat emprunteur à n'importe quel moment lors de la première année. Enfin, en 2017, l'amendement Bourquin permet aux assurés de substituer leur assurance annuellement à chaque date d'anniversaire.

Ces réformes devraient entraîner un changement de comportement des consommateurs vis-à-vis de l'assurance emprunteur. Cela pourrait se traduire notamment par une augmentation des taux de résiliation. Ce phénomène viendrait donc s'ajouter à celui du rachat anticipé de crédit, conséquence directe de la baisse importante des taux d'intérêts. Une augmentation des taux de résiliation serait alors synonyme de privation des flux futurs pour l'assureur et viendrait affecter directement la rentabilité des contrats.

L'essor du *Machine Learning* offre un grand nombre d'outils de modélisation permettant d'envisager de nouvelles manières de modéliser le risque de résiliation. Toutefois, rien n'indique que celles-ci soient suffisamment fiables pour justifier d'être employées au dépend des méthodes classiques. De plus, les méthodes de *Machine Learning* présentent des contraintes techniques dont la prise en compte est nécessaire dans le cadre d'une mise en application.

Ces problématiques sont d'autant plus intéressantes que nous appliquerons ces méthodes de *Machine Learning* à une base de données possédant une quantité d'informations limitée en comparaison des bases auxquelles celles-ci sont généralement appliquées. Ainsi, une confrontation, tant d'un point de vue technique que financier pour l'assureur, d'une méthode classique à des méthodes de *Machine Learning* nous permettra de statuer sur l'intérêt de ces dernières.

Pour répondre à cette question, nous serons amenés à confronter modélisation actuarielle et modélisation *Machine Learning*. Dans un premier chapitre, nous débuterons par une présentation des généralités relatives à l'assurance emprunteur et nous définirons le cadre de l'étude. Puis, dans un second chapitre, nous établirons des taux de résiliation à l'aide de l'estimateur de Kaplan-Meier et du lissage de Whittaker-Henderson. Ensuite, dans un troisième chapitre, nous construirons des taux de résiliation à l'aide de méthodes de *Machine Learning* telles que les arbres de décision, *Random Forest* et *XGBoost*. Enfin, dans un quatrième et dernier chapitre, nous éprouverons ces différentes méthodes en étudiant les impacts des différents taux obtenus sur la rentabilité du contrat.

Table des matières

Abréviations et acronymes

AEREAS	Assurer et emprunter avec un risque aggravé de santé
AT	Arrêt de travail
CI	Capital initial
CRD	Capital restant dû
CSR	Capital sous risque
DC	Décès
IPP	Incapacité permanente partielle
IPT	Incapacité permanente totale
ITT	Incapacité temporaire totale
PM	Provision mathématique
PRC	Provision pour risques croissants
PSAP	Provisions pour sinistres à payer
PTIA	Perte totale et irréversible d'autonomie
VAP	Valeur actuelle probable

Chapitre 1

L'assurance emprunteur

L'objectif de ce chapitre est de présenter dans un premier temps le marché de l'assurance emprunteur en France puis, dans un second temps, d'avoir une approche théorique des produits emprunteurs.

1.1 Le marché de l'assurance emprunteur

1.1.1 Caractéristiques du marché

Selon un rapport de l'Inspection Générale des Finances publié en 2013 [IGF, 2013], le contrat d'assurance souscrit dans le cadre d'un prêt représenterait entre 5% et 20% du coût total du crédit. Les assureurs s'appuyant sur le caractère obligatoire de l'assurance emprunteur pour pratiquer des tarifs nettement supérieurs au tarif couvrant le risque pur, le gouvernement a mis en place un certain nombre de réformes afin de faire diminuer ces marges au profit de tarifs plus compétitifs pour les consommateurs. Les chiffres présentés par la suite seront ceux correspondant à l'année 2018 rapportée par la convention AERAS [Ministère de l'Economie et des Finances, 2018].

Depuis plusieurs années, le marché de l'assurance emprunteur est en constante augmentation allant jusqu'à atteindre les 9,4 milliards d'euros de cotisations en 2018. Cette évolution peut en partie s'expliquer par la hausse importante du prix de l'immobilier dans les grandes villes de France ces 10 dernières années entraînant un accroissement des montants empruntés. Ainsi, le montant des cotisations s'élève mécaniquement.

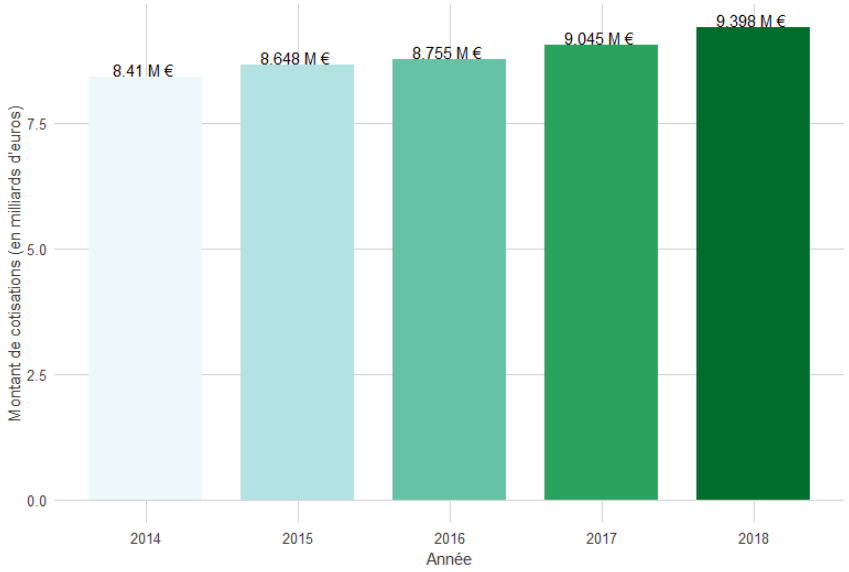


FIGURE 1.1 – Évolution du montant des cotisations en assurance emprunteur

Les 9,4 milliards d'euros de cotisations se répartissent de la manière suivante selon le type de prêts :

- 6 673 millions d'euros au titre de prêts immobiliers, soit 71% des cotisations.
- 2 068 millions d'euros au titre de prêts à la consommation, soit 22% des cotisations.
- 658 millions d'euros au titre de prêts professionnels, soit 7% des cotisations.

Ces chiffres montrent que le marché de l'assurance emprunteur est essentiellement tourné vers les particuliers. En effet, ces derniers représentent 93% des cotisations du secteur. Il s'agit donc du levier principal sur lequel le gouvernement peut agir s'il souhaite intervenir sur le marché.

Avant les différentes lois visant l'ouverture à la concurrence, le marché de l'assurance emprunteur était dominé par les organismes de bancassurance (Crédit Agricole, BPCE,...). Alors que l'ouverture à la concurrence aurait dû faire augmenter la part de cotisations en délégation d'assurance, nous observons une stabilité de celle-ci. En effet, entre 2014 et 2018, la part de cotisations en délégation est restée la même, aux alentours de 11,8%. La délégation d'assurance correspond au fait de faire assurer son prêt auprès d'un assureur autre que celui proposé par l'établissement de crédit. Cette stabilité peut s'expliquer notamment par des assurés qui ne souhaitent pas investir de temps pour effectuer les démarches de résiliation ou encore par le fait que les organismes bancaires ont actualisé leur tarification pour une segmentation plus importante reflétant mieux le risque couvert. Le manque de communication

auprès des particuliers autour de cette réforme peut aussi apporter un élément d'explication à ce phénomène.

Enfin, nous observons que deux types de garanties sont très majoritairement souscrites : la garantie Décès/PTIA et la garantie Incapacité/Invalidité. Dans une moindre mesure, la garantie perte d'emploi est elle aussi souscrite. La répartition des garanties en terme de volume de cotisations en 2018 est la suivante :

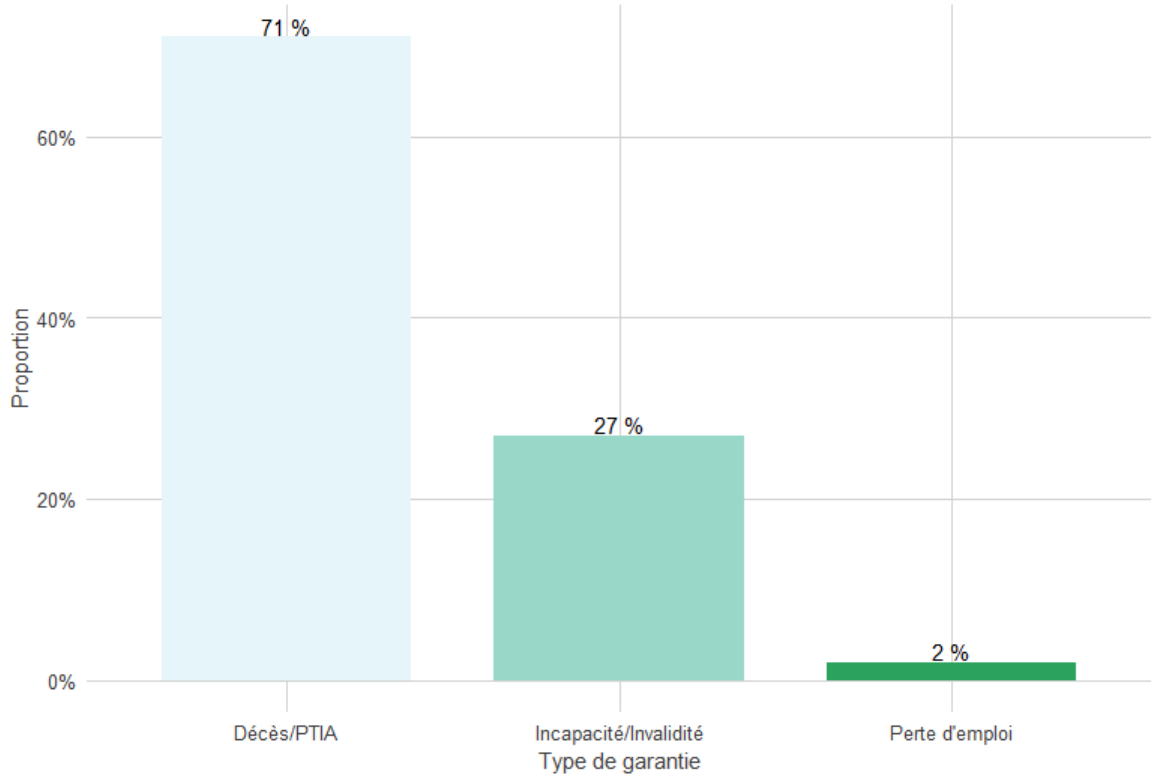


FIGURE 1.2 – Répartition des cotisations de 2018 par type de garantie

La prédominance de la garantie Décès/PTIA s'explique par le fait que celle-ci soit obligatoire sur tous les contrats souscrits. Ce n'est pas le cas pour les autres garanties qui sont optionnelles.

1.1.2 Le profil de l'emprunteur pour le crédit immobilier

Selon une étude menée par France Info en 2019 [France Info, 2019], le profil type de l'emprunteur en 2018 était une personne de 36 ans qui empruntait environ 217 k € sur une durée de 20 ans. Avec un coefficient de corrélation de - 0,71 calculé sur la période de 2001 à 2020, nous pouvons observer que l'âge à la souscription a tendance à baisser quand la durée d'emprunt a tendance à augmenter [Meilleurtaux.com, 2019].

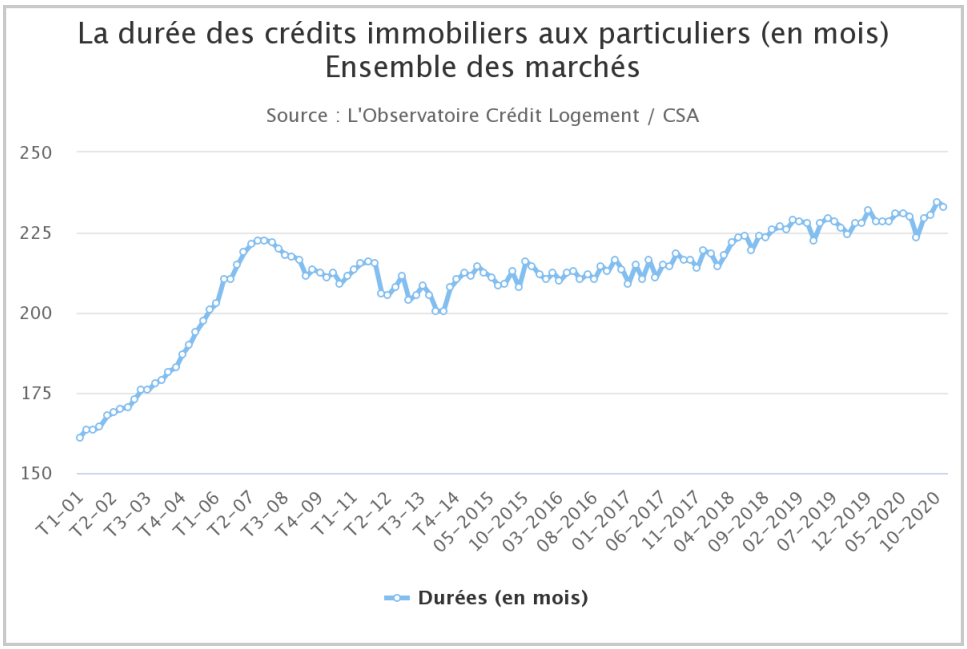


FIGURE 1.3 – Evolution de la durée d'emprunt depuis 2001

Cette augmentation de la durée d'emprunt s'explique notamment par la baisse constante des taux moyen d'emprunt depuis quelques années comme en atteste le graphe suivant issu de l'Observatoire Crédit Logement :

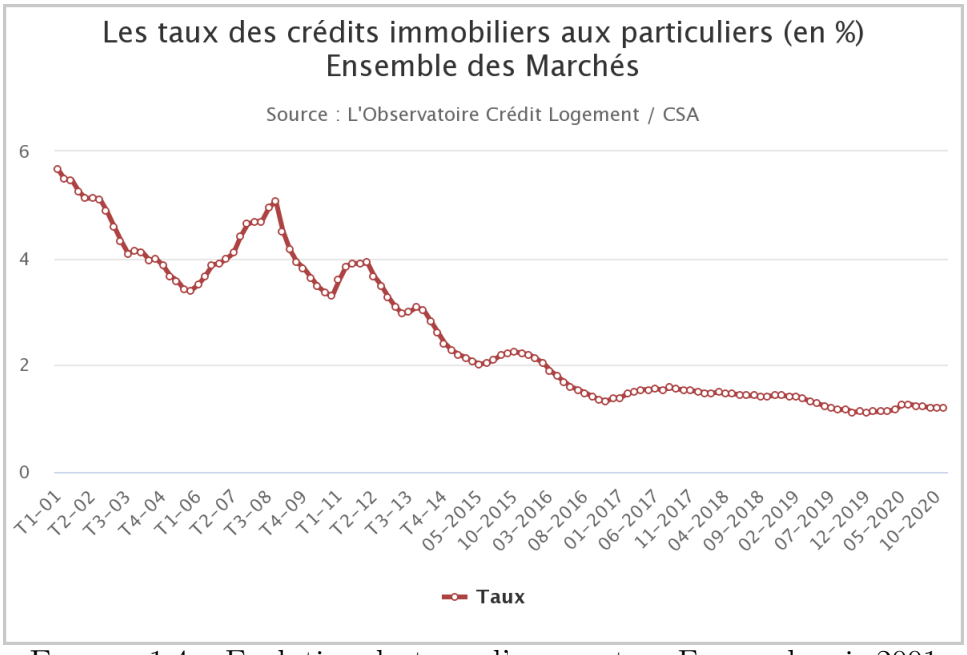


FIGURE 1.4 – Evolution du taux d'emprunt en France depuis 2001

Les économistes ne prévoyant pas de hausse des taux d'emprunt dans un futur proche, cette situation tend à se maintenir, renforçant ainsi la nécessité d'appréhender au mieux les risques liés à ce marché et notamment les risques liés à la résiliation.

1.1.3 Évolutions réglementaires

Les évolutions réglementaires qu'a connue l'assurance emprunteur suivent toutes le même objectif : diminuer les marges réalisées par les organismes bancaires en favorisant la concurrence afin que l'assurance emprunteur présente un coût moindre pour le consommateur tout en le protégeant. Nous pouvons noter quatre évolutions réglementaires majeures au cours des quinze dernières années [1].

1.1.3.1 Convention AERAS - 2007

Entrée en vigueur en 2007, la convention AERAS, acronyme de "s'Assurer et Emprunter avec un Risque Aggravé de Santé " a pour objectif de permettre notamment aux personnes ayant des antécédents médicaux importants d'accéder au système d'emprunt bancaire via une étude approfondie des risques et une limitation de la majoration des tarifs. Cette convention impacte donc directement le marché de l'assurance emprunteur au travers des contrats souscrits via les prêts accordés par la convention AERAS. En 2015, le "droit à l'oubli" est ajouté à cette convention et permet l'élargissement de la liste des maladies prises en compte par la convention AERAS. De plus, le droit à l'oubli permet aux personnes qui n'ont pas eu de rechute depuis un grand nombre d'années d'être exemptées de déclaration médicale et par conséquent de ne plus être sujet à des exclusions de garantie ni à des majorations tarifaires.

1.1.3.2 Loi Lagarde - 2010

La loi Lagarde, aussi appelée Réforme du crédit à la consommation, est la première réforme d'une série de trois lois visant à délier le prêt bancaire de son assurance. L'élaboration de cette loi commence en 2008 pour finalement entrer en vigueur en 2010. Elle oblige les banques à fournir une fiche d'information standardisée résumant les détails de l'offre de prêt. De plus, elle donne l'opportunité à l'assuré de choisir son assurance emprunteur : à partir du moment où il présente une offre aux garanties équivalentes, la banque est obligée d'accepter l'assurance et d'accorder le prêt sans en modifier ni le taux ni ajouter des frais de délégation.

1.1.3.3 Loi Hamon - 2014

En 2014, la loi relative à la consommation - dite loi Hamon - intègre un article ayant pour objectif d'alléger la procédure de résiliation des contrats d'assurance emprunteur. Désormais, l'emprunteur a la possibilité de résilier son assurance à n'importe quel moment pendant un an à compter de la date de signature du contrat. Les nouvelles garanties doivent cependant être au moins équivalentes à celles du contrat résilié.

1. Article L313-30 du Code de la Consommation

1.1.3.4 Amendement Bourquin - 2017

Enfin, en 2017, le projet de loi relatif à la transparence, à la lutte contre la corruption et à la modernisation de la vie économique, plus communément appelé loi Sapin 2 ou Amendement Bourquin, renforce encore plus les mesures mentionnées précédemment. En effet, depuis son entrée en vigueur, un emprunteur peut exercer son droit à la résiliation de son contrat d'assurance emprunteur à chaque date d'anniversaire du contrat.

1.1.4 Enjeux

Dans un contexte de quasi-monopole des organismes de bancassurance, les évolutions réglementaires ont pour but d'ouvrir encore plus ce marché à la concurrence. Cette nouvelle composante de résiliation vient s'ajouter au phénomène de rachat de crédit déjà présent. Par conséquent, une détermination des taux de résiliation est nécessaire afin d'ajuster les paramètres des produits présents sur le marché.

Il est indispensable de faire la distinction entre taux de rachat et taux de résiliation. En effet, le rachat correspond au remboursement anticipé des dernières échéances du prêt, ce remboursement implique la fin du contrat car le montant emprunté est totalement remboursé. La résiliation d'un contrat fait quant à elle référence à des motifs plus larges. Cette dernière peut provenir du rachat d'un prêt mais peut aussi venir du choix d'un assuré de changer de contrat d'assurance emprunteur en passant à la délégation d'assurance par exemple. Dans ce cas, l'assuré a toujours son obligation de remboursement du prêt à chaque échéance mais l'organisme qui l'assure a changé.

La méthode utilisée pour calculer ces taux de résiliation est donc capitale afin d'estimer au mieux la réalité du portefeuille. Dans un contexte de multiplication des méthodes de calcul, notamment du fait de l'évolution de la Data Science, il est donc intéressant d'évaluer la pertinence de ces méthodes dans le cadre de l'assurance emprunteur.

1.2 Théorie de l'assurance emprunteur

Cette partie a pour objectif d'établir une présentation théorique de l'assurance emprunteur. Bien que plus large que la problématique d'estimation des taux de résiliation, cette thématique est néanmoins nécessaire afin de comprendre dans quel contexte s'inscrit l'étude réalisée.

1.2.1 Principe de fonctionnement

Contrairement à un contrat d'assurance classique, l'assurance emprunteur est une assurance qui lie trois parties différentes :

- l'établissement de crédit ;
- l'assureur ;
- l'assuré.

Elle garantit à l'établissement de crédit le remboursement du crédit en cas de défaut de l'assuré pour une cause prévue au contrat de l'assurance. Elle permet également de protéger l'assuré en garantissant la conservation des biens et la non-transmission des dettes aux héritiers en cas de décès ou assimilé.

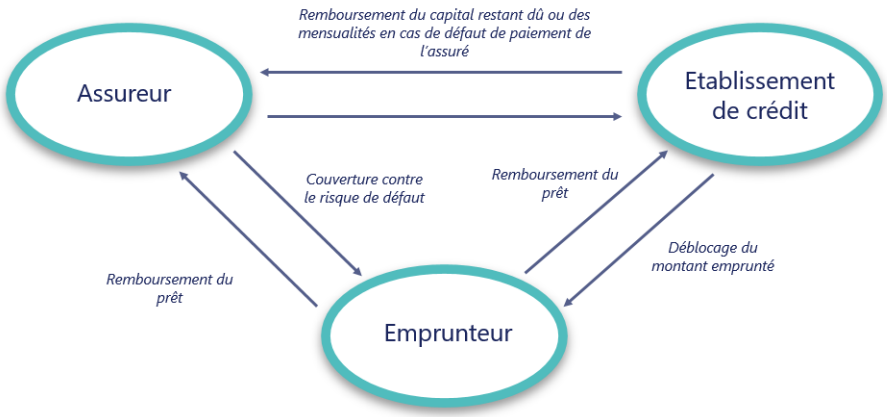


FIGURE 1.5 – Fonctionnement de l'assurance emprunteur

D'un point de vue légal, le Code de la Consommation [2] régit les conditions générales du crédit et le Code des Assurances [3] régit les conditions particulières des contrats d'assurance emprunteur.

2. Art. L311-1 et L313-40 à L313-45
3. Art. L113-1 à L113-17

Ainsi, les quatre parties classiques du contrat d'assurance sont représentées par les entités suivantes :

- **l'assuré :** personne qui emprunte de l'argent auprès de l'établissement de crédit ;
- **l'assureur :** compagnie d'assurance qui s'engage à rembourser le capital restant dû en cas de défaut de l'assuré ;
- **le bénéficiaire :** établissement de crédit ayant accordé le prêt ;
- **le souscripteur :** personne physique ou morale qui souscrit le contrat. Dans le cas d'un contrat collectif c'est l'organisme bancaire, dans le cas d'un contrat individuel, c'est l'assuré.

En théorie, l'assurance emprunteur ne revêt pas de caractère obligatoire. Cependant, en pratique, les organismes prêteurs exigent quasi systématiquement la souscription d'un contrat emprunteur. Il est alors possible de faire la distinction entre deux types de contrats :

- **le contrat collectif :** définit à l'article L141-1 du Code des Assurances, c'est le type de contrat que proposent les banques à leurs clients. Dans ce type de contrat, le souscripteur est une personne morale en la présence de la banque. Ce contrat dispose généralement d'une faible segmentation, appliquant ainsi un taux de prime similaire pour tous les assurés d'une même classe d'âge ;
- **le contrat individuel :** depuis 2010, avec la promulgation de la loi Lagarde, un assuré est en droit de contracter une assurance emprunteur à titre personnel sans se voir opposer de refus de la banque à la condition que cette assurance présente des garanties au minimum identiques au contrat proposé par la banque. Ce type de contrat est souvent très fortement segmenté ce qui le rend très attractif pour les individus peu risqués à savoir les jeunes ;

1.2.2 Garanties proposées

De nombreuses garanties sont proposées dans un contrat emprunteur. Cependant, toutes ne sont pas rattachées aux mêmes branches au sens du Code des Assurances. On peut séparer les différentes garanties en deux catégories :

- **La garantie vie :** classifiée dans la branche 20 de l'article R321-3 du Code des Assurances, elle assure la prise en charge du remboursement du prêt en cas de décès de l'assuré ;

– **Les garanties non-vie :** classifiées dans les branches 1 et 2 de l'article R321-3 du Code des assurances, elles assurent la prise en charge des mensualités en cas d'arrêt de travail de l'assuré.

Parmi ces garanties, toutes ne sont pas obligatoires mais certaines se retrouvent dans presque tous les contrats. La garantie décès est présente dans tout contrat d'assurance emprunteur. En cas de survenance du décès de l'emprunteur, l'assurance verse le capital sous risque restant dû et les intérêts courus entre la dernière échéance et le décès à l'organisme bancaire. Le capital sous risque restant dû s'obtient en calculant le produit du capital restant dû et de la quotité, part du capital couvert par la garantie.

La garantie Perte Totale et Irréversible d'Autonomie, aussi appelée garantie PTIA, couvre les cas où l'assuré se retrouve dans l'obligation de faire intervenir une tierce personne pour l'aider dans ses tâches de la vie quotidienne. Cette garantie est obligatoire dans un contrat emprunteur jusqu'à 65 ans. Une fois cet âge dépassé ou au moment du passage à la retraite, la garantie devient nulle. L'activation de la garantie PTIA donne généralement lieu au remboursement des mensualités par l'assureur mais, dans certains cas, le remboursement du capital sous risque restant dû est proposé lorsqu'il est certain que l'état de santé de l'assuré ne s'améliorera plus.

Les garanties incapacité parmi lesquelles nous retrouvons la garantie Incapacité Temporaire Partielle (ITP) et la garantie Incapacité Temporaire Totale (ITT). Ces deux garanties concernent les cas où l'assuré n'est plus, de manière temporaire, en mesure d'exercer son activité professionnelle à temps complet. Ces garanties s'accompagnent généralement d'un délai de franchise incompressible durant lequel l'assuré ne peut pas prétendre au versement de ses prestations. Dans ce cas, l'assurance prend le relais sur une partie du remboursement des échéances pouvant aller jusqu'à une prise en charge totale dans le cas de l'incapacité temporaire totale. Le caractère temporaire de cette couverture justifie l'existence de la garantie qui va suivre.

Les garanties invalidité parmi lesquelles nous retrouvons la garantie Invalidité Permanente Partielle (IPP) et la garantie Invalidité Permanente Totale (IPT). Contrairement aux garanties incapacité, ces garanties couvrent le cas où l'assuré se retrouve dans l'impossibilité physique ou mentale définitive d'exercer une activité professionnelle à temps plein. Dans ce cas, l'assurance prend en charge un pourcentage, définit contractuellement, des remboursements futurs sans contrainte de temps.
Il existe également des garanties qui sont beaucoup moins présentes dans les contrats emprunteurs. Parmi elles, nous trouvons notamment la garantie perte d'emploi, rattachée à

la branche 16 du Code des Assurances, qui vient pallier la perte de revenus liée à une période de chômage. Nous pourrions également nommer les garanties décès accidentel, perte de ressources ou encore protection revente.

1.2.3 Types de prêts

L'assurance emprunteur concerne la couverture de plusieurs types de prêts. Nous explicitons ici leur nature ainsi que le poids qu'ils occupent dans l'économie de l'assurance emprunteur :

- **Le prêt immobilier :** il permet de financer l'achat d'un bien immobilier. Il représente souvent un montant emprunté élevé et s'étale sur un grand nombre d'années. C'est le type de prêt que l'on retrouve majoritairement dans le monde de l'assurance emprunteur avec 71% des cotisations en 2018 ;
- **Le prêt à la consommation :** il permet de financer des gros achats de la vie quotidienne (automobile, mobiliers, ...) ou bien permet de régler des factures imprévues. Les montants empruntés et la durée sont plus faibles que pour le prêt immobilier. En 2018, le prêt à la consommation représente 22% des cotisations de l'assurance emprunteur ;
- **Le prêt professionnel :** il permet aux entreprises de financer l'acquisition de biens (machines, matériel, immobiliers ...). Ce type de prêt occupe une place peu importante dans le cadre de l'assurance emprunteur avec seulement 7% des cotisations en 2018.

1.2.4 Modalités de remboursement

En assurance emprunteur, les prestations versées par la compagnie d'assurance dépendent du capital restant dû sous risque en cas de décès de l'assuré et des mensualités payées en cas de passage en incapacité/invalidité ou d'activation de la garantie PTIA. Or, la manière dont est remboursé le crédit a une grande importance sur les montants mentionnés précédemment. Il est donc important d'y apporter une attention particulière. Nous introduirons ici les notations générales du remboursement de crédit et nous détaillerons les différentes manières de le rembourser. Dans la suite, nous considérerons une quotité de 100 %. Ainsi, le capital restant dû sous risque sera égal au capital restant dû.

Dans le cadre du remboursement d'un emprunt et de la tarification du contrat d'assurance emprunteur qui lui est associé, les notations communément utilisées sont les suivantes :

- E : Montant emprunté ;
- j : le taux d'intérêt de l'emprunt ;

- n : la durée du prêt ;
- A_k : le montant du capital amorti à l'échéance k ;
- CRD_k : le capital restant dû à l'échéance k ;
- I_k : intérêts versés à l'échéance k ;
- R_k : remboursement versé à l'échéance k ;
- $EA_{(g)}$: l'engagement de l'assureur pour la garantie étudiée. Ici l'indice g correspond à une garantie décès (DC) ou une garantie arrêt de travail (AT) ;
- $Ea_{(g)}^{(l)}$: l'engagement de l'assuré pour la garantie étudiée. Ici l'indice l correspond à une tarification au capital initial (CI) ou une tarification au capital restant dû (CRD) ;
- x : âge de l'assuré à la souscription du contrat. Nous ne considérons que des valeurs entières ;
- i : le taux d'actualisation annuel. Il correspond au taux technique vie en vigueur au moment de la souscription du contrat ;
- $i^{(m)} = (1+i)^{\frac{1}{12}} - 1$: le taux d'actualisation mensuel calculé à partir du taux d'actualisation annuel ;
- $v = \frac{1}{1+i}$ et $v^{(m)} = \frac{1}{1+i^{(m)}}$: les facteurs d'actualisation respectivement annuel et mensuel ;
- ${}_kp_x$: probabilité pour un individu d'âge x de rester en vie les k années suivantes ;
- q_{x+k} : probabilité d'un individu d'âge $x+k$ de mourir dans l'année qui suit ;
- ${}_kr_{x,x+t}$: probabilité pour un individu d'âge à la souscription x et d'âge atteint $x+t$ de ne pas racheter son emprunt entre les anciennetés t et $k+t$;
- $w = min(n-t; x_{AT}^{max} - x - t)$: correspond au nombre d'années pendant lesquelles l'assuré est couvert par la garantie arrêt de travail. Si celui-ci part en retraite avant la fin de son prêt il est possible qu'il ne soit plus couvert par cette garantie ;
- $e_x^{(12)}$: la probabilité d'entrée en arrêt de travail à l'âge x ;
- m_{AT} : interpolation linéaire entre les probabilités de maintien en arrêt de travail à l'âge x. Nous l'obtenons de la manière suivante :

$$m_{AT}(x) = \frac{1}{2} \sum_{j=0}^{12*w-1} (v^{(12)})^j * {}_jp_x^{AT} + (v^{(12)})^{j+1} * {}_{j+1}p_x^{AT}.$$

et ${}_jp_x^{AT}$ correspond à la probabilité de maintien en arrêt de travail d'une personne d'âge x pendant j mois ;

– $\lfloor . \rfloor$ correspond à l'opérateur de la partie entière. Ici, $\lfloor \frac{k}{12} \rfloor$ correspond donc à l'année en cours au moment k.

Nous pouvons d'ores et déjà établir des relations entre les différentes valeurs mentionnées précédemment. Ces relations seront valables pour toutes les échéances comprises entre le début et la fin de l'emprunt :

– $E = CRD_0$: en $k = 0$, le capital restant dû est égal au montant emprunté ;

– $A_k = CRD_{k-1} - CRD_k$: l'amortissement du capital effectué à l'échéance k est égal à la part du capital remboursé à cette échéance ;

– $I_k = CRD_{k-1} * j$: les intérêts versés à l'échéance k sont proportionnels au capital restant dû à l'échéance $k - 1$;

– $R_k = A_k + I_K$: le montant versé à l'échéance k correspond à l'amortissement du capital et aux intérêts de l'échéance k ;

Il reste alors à définir une formule itérative permettant de calculer le capital restant dû à chaque échéance k. Grâce aux différentes relations obtenues ci-dessus, nous obtenons la formule suivante :

$$\begin{aligned} R_k &= A_k + I_k \\ R_k &= (CRD_{k-1} - CRD_k) + CRD_{k-1} * j \\ R_k &= CRD_{k-1} * (1 + j) - CRD_k \end{aligned}$$

Par conséquent, nous avons finalement :

$$CRD_k = CRD_{k-1} * (1 + j) - R_k.$$

1.2.4.1 Remboursement à amortissement constant

Dans le cadre du remboursement d'un prêt à amortissement constant, la part du capital remboursée est la même tous les ans. Le terme de remboursement linéaire du crédit est également utilisé. Nous avons alors pour toutes les échéances k la formule suivante, avec n non nul :

$$A_k = A = \frac{E}{n}.$$

Alors, par récurrence, la formule explicite du capital restant dû à l'échéance k est la suivante :

$$CRD_k = E * \frac{n-k}{n}.$$

Afin de mieux visualiser les différents montants versés à chaque échéance, un tableau d'amortissement peut être utilisé pour observer le déroulement de la vie du prêt. Dans le cadre d'un crédit de 100 000€ au taux nominal de 1% sur 10 ans remboursé annuellement, le tableau d'amortissement peut être présenté de la manière suivante :

Année	**Capital restant dû**	**Intérêt**	**Amortissement**	**Remboursement**
0	100 000 €	0 €	0 €	0 €
1	90 000 €	1 000 €	10 000 €	11 000 €
2	80 000 €	900 €	10 000 €	10 900 €
3	70 000 €	800 €	10 000 €	10 800 €
4	60 000 €	700 €	10 000 €	10 700 €
5	50 000 €	600 €	10 000 €	10 600 €
6	40 000 €	500 €	10 000 €	10 500 €
7	30 000 €	400 €	10 000 €	10 400 €
8	20 000 €	300 €	10 000 €	10 300 €
9	10 000 €	200 €	10 000 €	10 200 €
10	0 €	100 €	10 000 €	10 100 €
Total	-	**5 500 €**	**100 000 €**	**105 500 €**

TABLE 1.1 – Tableau d'amortissement d'un prêt à amortissement constant

1.2.4.2 Remboursement à annuités constantes

Dans ce deuxième cas, c'est le montant remboursé à chaque échéance qui est constant. Nous nous retrouvons alors avec la formule suivante, pour j non nul :

$$R_k = R = E * \frac{j}{1-(1+j)^{-n}}.$$

Pour un remboursement à annuités constantes, la part de capital amortie à chaque échéance

est croissante dans le temps. Les intérêts sont principalement payés sur les premières annuités et le capital est remboursé sur les dernières annuités comme en atteste le tableau d'amortissement suivant :

Année	Capital restant dû	Intérêt	Amortissement	Remboursement
0	100 000,00 €	0 €	0 €	0 €
1	90 441,79 €	1 000,00 €	9 558,21 €	10 558,21 €
2	80 788,00 €	904,42 €	9 653,79 €	10 558,21 €
3	71 037,67 €	807,88 €	9 750,33 €	10 558,21 €
4	61 189,84 €	710,38 €	9 847,83 €	10 558,21 €
5	51 243,53 €	611,90 €	9 946,31 €	10 558,21 €
6	41 197,76 €	512,44 €	10 045,77 €	10 558,21 €
7	31 051,53 €	411,98 €	10 146,23 €	10 558,21 €
8	20 803,84 €	310,52 €	10 247,69 €	10 558,21 €
9	10 453,67 €	208,04 €	10 350,17 €	10 558,21 €
10	0 €	104,54 €	10 453,67 €	10 558,21 €
Total	-	**5 582,10 €**	**100 000 €**	**105 582,10 €**

TABLE 1.2 – Tableau d'amortissement d'un prêt à annuités constantes

1.2.4.3 Remboursement *In Fine*

Dans le cas d'un remboursement de crédit In Fine, le capital est remboursé en une seule fois à la dernière échéance. Pour toutes les autres échéances, l'emprunteur ne paye que des intérêts qui seront fixes d'une échéance à l'autre. Ce type de remboursement permet aux personnes qui ont la possibilité de rembourser leur capital en une seule fois de bénéficier d'avantages fiscaux, notamment au niveau de la déduction des intérêts des revenus locatifs.

1.2.5 Tarification et provisionnement

Dans cette partie, nous traiterons en détail les deux grandes méthodes de tarification en assurance emprunteur ainsi que les provisions que l'assureur se doit de constituer pour faire face à ses engagements.

1.2.5.1 Prime au capital initial ou au capital restant dû

En assurance emprunteur, deux grands principes de tarification se distinguent : la tarification au **capital initial** et la tarification au **capital restant dû**.

La tarification au capital initial est certainement la plus répandue. Elle consiste en l'application d'un taux de prime fixe au capital emprunté. La prime payée par l'assuré est alors constante tout au long du remboursement du prêt. Cette méthode de tarification présente l'avantage d'être facile à vendre commercialement, les assurés payent le même tarif à chaque échéance et peuvent donc mieux appréhender leur échéancier. Néanmoins, ce type de tarification est généralement peu segmenté et s'applique souvent à la tarification des contrats collectifs.

Le principe peut être résumé selon la formule suivante :

$$Prime = Taux\ de\ prime * Capital\ emprunté$$

La tarification au capital restant dû consiste, quant à elle, en un pourcentage du capital restant dû à chaque échéance. Contrairement à la tarification au capital initial, le taux de prime évolue chaque année. La combinaison de ces deux variables entraîne le fait que la prime payée par l'assuré ne va pas être constante dans le temps. Ce type de tarification est caractéristique des contrats individuels car souvent très dépendants de caractéristiques propres à l'assuré comme son âge ou s'il est fumeur par exemple. Par ailleurs, le principe de tarification au capital restant dû peut être résumé avec la formule suivante :

$$Prime_k = Taux\ de\ prime_k * Capital\ Restant\ D\hat{u}_k$$

Dans les deux cas, la tarification consiste en l'égalisation des engagements de l'assureur et de l'assuré. Pour la tarification au capital initial, elle se fait une seule fois au moment de la souscription du contrat. Pour la tarification au capital restant dû, elle se fait à chaque échéance k.

1.2.5.2 Tables réglementaires

En ce qui concerne la tarification du risque décès en assurance emprunteur, le Code des Assurances préconise l'utilisation des tables réglementaires homologuées par le ministère de l'économie et des finances [4]. Il est toutefois possible de leur appliquer un coefficient d'abattement pour mieux correspondre au portefeuille étudié ou bien d'utiliser une table d'expérience certifiée.

4. Article A.132-18 du Code des Assurances

Pour ce qui est de la tarification du risque arrêt de travail, nous pouvons procéder de différentes manières. Une première approche consiste à utiliser une table mixte qui regroupe à la fois le risque de passage en incapacité et le risque de passage en invalidité. Une seconde approche consiste à différencier ces deux risques en utilisant deux tables de passages distinctes. Dans les deux cas, il est nécessaire d'utiliser une table de maintien de l'état dans lequel est l'assuré. Enfin, une approche supplémentaire consiste à différencier les probabilités de décès de l'assuré, en fonction de l'état dans lequel il se trouve, au lieu de considérer une probabilité unique quelque soit l'état de l'assuré. La probabilité de décès d'un individu dans l'année sera alors différente selon qu'il est en bonne santé, en incapacité ou en invalidité.

Dans la suite de ce chapitre, les probabilités utilisées seront donc celles calculées conformément aux tables préconisées par le code des assurances à savoir la table TH 00-02 pour le risque décès et la table de maintien du BCAC pour le risque arrêt de travail.

1.2.5.3 Tarification du risque décès

Pour une garantie décès l'assureur s'engage, sur toute la durée du contrat, à verser l'intégralité du capital restant dû sous risque à l'organisme prêteur en cas de survenance du décès de l'assuré. Dans notre cas, nous considérerons que le remboursement du prêt est mensuel et que le décès intervient systématiquement en milieu d'année. Si nous prenons en compte un pas annuel, nous nous retrouvons alors, pour tout $t \in [\![0; n-1]\!]$, avec l'engagement de l'assureur suivant :

$$EA_{DC}(t) = \sum_{k=0}^{n-t-1} CRD_{12*(t+k)+6} * v^{k+\frac{6}{12}} * {}_kp_{x+t} * q_{x+t+k} * {}_kr_{x,x+t}.$$

En théorie, dans le Code des Assurances, l'engagement de l'assureur n'intègre pas la composante de résiliation du contrat. Toutefois, pour une tarification au capital initial, l'ajout de ce terme permet d'avoir une meilleure approche de l'engagement réel de l'assureur. Par conséquent, pour une tarification au capital restant dû, l'engagement de l'assureur est le même en retirant le terme ${}_kr_{x,x+t}$.

Tarification au capital initial

Dans le cadre d'une tarification au capital initial, l'engagement de l'assuré est le suivant :

$$Ea_{DC}^{(CI)}(t) = \text{taux de prime}_{DC}^{(CI)} * E * \sum_{k=0}^{n-t-1} v^k * {}_kp_{x+t} * {}_kr_{x,x+t}.$$

Pour retrouver le taux de prime utilisé dans le cadre de la tarification du risque décès en capital initial, nous appliquons alors le principe d'équité actuarielle, c'est-à-dire que l'on égalise les engagements assureur et assuré, au moment de la souscription du contrat, à savoir en $t = 0$. Par conséquent, le taux de prime s'obtient de la manière suivante :

$$\begin{aligned} EA_{DC}(0) &= Ea_{DC}^{(CI)}(0) \\ \sum_{k=0}^{n-1} CRD_{12*k+6} * v^{k+\frac{6}{12}} * {}_kp_x * q_{x+k} * {}_kr_x &= \text{taux de prime}_{DC}^{(CI)} * E * \sum_{k=0}^{n-1} v^k * {}_kp_x * {}_kr_x \\ \text{taux de prime}_{DC}^{(CI)} &= \frac{\sum_{k=0}^{n-1} CRD_{12*k+6} * v^{k+\frac{6}{12}} * {}_kp_x * q_{x+k} * {}_kr_x}{E * \sum_{k=0}^{n-1} v^k * {}_kp_x * {}_kr_x} \end{aligned}$$

.

De plus, $\forall (k, x), v^k * {}_kp_x > 0$

Tarification au capital restant dû

Dans le cas de la tarification au capital restant dû, le taux de prime appliqué change chaque année. Nous nous retrouvons alors, pour tout $t \in [\![0; n-1]\!]$, avec l'engagement assuré suivant :

$$Ea_{DC}^{(CRD)}(t) = \sum_{k=0}^{n-t-1} \text{taux de prime}_{DC,t+k}^{(CRD)} * CRD_{12*(t+k)} * v^k * {}_kp_{x+t}.$$

Ainsi, pour obtenir les différents taux de prime, nous égalisons les engagements assureur et assuré chaque année. Nous nous retrouvons alors pour tout $k \in [\![0; n-1]\!]$ avec le taux de prime suivant :

$$\begin{aligned} EA_{DC,k} &= Ea_{DC,k}^{(CRD)} \\ CRD_{12*k+6} * v^{k+\frac{6}{12}} * {}_kp_x * q_{x+k} &= \text{taux de prime}_{DC,k}^{(CRD)} * CRD_{12*k} * v^k * {}_kp_x \\ \text{taux de prime}_{DC,k}^{(CRD)} &= \frac{CRD_{12*k+6} * v^{k+\frac{6}{12}} * {}_kp_x * q_{x+k}}{CRD_{12*k} * v^k * {}_kp_x} \\ \text{taux de prime}_{DC,k}^{(CRD)} &= \frac{CRD_{12*k+6}}{CRD_{12*k}} * v^{\frac{6}{12}} * q_{x+k}. \end{aligned}$$

De plus, $\forall k, CRD_{12*k} > 0$

1.2.5.4 Tarification du risque arrêt de travail

Dans le cadre de notre étude, nous conserverons l'hypothèse d'utilisation d'une table mixte pour prendre en compte le passage en incapacité ou en invalidité. De plus, dans le cadre de la garantie arrêt de travail, l'assurance ne rembourse pas le capital restant dû à l'organisme de crédit mais vient se substituer à l'assuré pour le remboursement des échéances, le temps que ce dernier reprenne le travail.

Contrairement à la garantie décès, nous considérons ici un pas mensuel étant donné que l'assureur sera amené à verser les mensualités à la place de l'assuré. Par conséquent, pour tout $t \in [\![0; n-1]\!]$, nous obtenons l'engagement de l'assuré suivant :

$$EA_{AT}(t) = \sum_{k=0}^{12*w-1} R_{12*t+k} * {}_kp_{x+t}^{(12)} * (v^{(12)})^k * e_{x+t+\lfloor \frac{k}{12} \rfloor}^{(12)} * m_{AT}(x+t+\lfloor \frac{k}{12} \rfloor) * {}_kr_{x,x+t}.$$

Comme pour la garantie décès, selon le Code des Assurances, la composante de résiliation n'est pas prise en compte dans le calcul des engagements de l'assureur. Ici aussi, l'ajout de ce terme dans une tarification au capital initial permet une meilleure estimation de l'engagement de ce dernier. L'engagement de l'assuré, quant à lui, ne change pas par rapport à la tarification du risque décès. Par conséquent, nous nous retrouvons ici aussi à différencier tarification au capital initial et tarification au capital restant dû.

Tarification au capital initial

En égalisant les engagements de l'assureur et de l'assuré au moment de la souscription du contrat, nous obtenons les équivalences suivantes :

$$\begin{aligned} EA_{AT}(0) &= Ea_{AT}^{(CI)} \\ EA_{AT}(0) &= \text{taux de prime}_{AT}^{(CI)} * E * \sum_{k=0}^{n-t-1} v^k * {}_kp_{x+t} * {}_kr_x \\ \text{taux de prime}_{AT} &= \frac{EA_{AT}(0)}{E * \sum_{k=0}^{n-t-1} v^k * {}_kp_{x+t} * {}_kr_x}. \end{aligned}$$

De plus, $\forall (k, x, t), v^k * {}_kp_{x+t} > 0$

Tarification au capital restant dû

Pour rappel, lorsque l'on tarifie au capital restant dû, il faut égaliser les engagements assureur et assuré pour obtenir chaque valeur du taux de prime. Dans le cadre de notre étude, le

taux de prime change tous les ans. Nous avons déjà calculé précédemment les engagements de l'assureur et de l'assuré. Nous obtenons donc les équivalences suivantes, pour tout $k \in [\![0; n-1]\!]$:

$$\begin{aligned} EA_{AT,k} &= Ea_{AT,k} \\ EA_{AT,k} &= \text{taux de prime}_{AT,k}^{(CRD)} * CRD_{12*k} * v^k * {}_kp_x \\ \text{taux de prime}_{AT,k}^{(CRD)} &= \frac{EA_{AT,k}}{CRD_{12*k} * v^k * {}_kp_x}. \end{aligned}$$

De plus, $\forall (k,x), CRD_{12*k} * v^k * {}_kp_x > 0$

1.2.5.5 Provisionnement

Lors de la vie d'une police d'assurance, la prime payée par l'assuré ne correspond pas toujours exactement au risque porté par l'assureur. Par exemple, dans le cas d'une prime nivelée, la prime payée par l'assuré au début du contrat est souvent plus faible que le risque réellement supporté par l'assureur à cet instant. Celui-ci est donc dans l'obligation légale de constituer une provision afin de pouvoir faire face à ses engagements en cas de réalisation du risque couvert.

Dans le cadre de l'assurance emprunteur, nous pouvons distinguer deux provisions principales :

- **Les provisions mathématiques (PM) :** définies à l'article R343-3 du Code des Assurances, ces provisions compensent, pour les opérations d'assurance sur la vie, l'écart entre les valeurs actuelles des engagements de l'assureur et de l'assuré.
- **Les provisions pour risque croissant (PRC) :** définies à l'alinéa 5 de l'article R343-7 du Code des Assurances, ces provisions compensent, pour les opérations d'assurance contre les risques de maladie et d'invalidité, l'écart entre les valeurs actuelles des engagements de l'assureur et de l'assuré.

Ces provisions sont calculées par assuré et par garantie. Nous avons ainsi les formules suivantes :

$$\begin{cases} PM_{DC}(t) = max\Big(EA_{DC}(t) - Ea_{DC}^{(l)}(t); 0\Big) \\ PRC_{AT}(t) = max\Big(EA_{AT}(t) - Ea_{AT}^{(l)}(t); 0\Big) \end{cases}$$

1.2.5.6 Cas des provisions pour sinistres à payer

Dans le cadre de la survenance d'un sinistre (décès par exemple), il est possible que l'intégralité du capital restant dû ne soit pas reversée à l'organisme de crédit en une seule fois. Dans cette hypothèse, l'assureur se doit de constituer une provision pour sinistre à payer (PSAP) correspondant au montant restant à payer sur le prochain exercice. Par exemple, dans le cas où la cadence de versement est de 70%/20%/10% et que l'assureur n'a versé que 70% de la somme, il se doit de constituer une provision pour sinistres à payer correspondant à 30% du capital restant dû.

Chapitre 2

Modélisation Actuarielle

Ce chapitre a pour but de détailler les méthodes actuarielles utilisées pour calculer les taux de résiliation du portefeuille considéré dans ce mémoire. Dans un premier temps, nous établirons une étude approfondie de la base de données utilisée pour l'étude. Dans un second temps, nous détaillerons les méthodes employées ainsi que les résultats obtenus.

2.1 Présentation de la base de données

Dans cette partie, nous allons présenter la base de données mise à notre disposition pour cette étude. Il s'agira de déterminer si le portefeuille est proche de ce qu'il est possible d'observer sur le marché précédemment présenté.

2.1.1 Structure de la base de données

2.1.1.1 Provenance des données brutes

Le produit emprunteur mis à notre disposition pour cette étude est un contrat d'assurance groupe commercialisé par un réseau de courtiers. Afin de répondre aux exigences de confidentialité des données, les informations concernant le nom et la localisation des courtiers commercialisant ce produit, ainsi que les nom et prénom des assurés seront anonymisés par la suite.

Ainsi, pour cette étude, la base de données est un ensemble de *reporting* mensuels s'étalant sur une période de sept ans de Juin 2011 à Mars 2017. Il est alors possible d'assurer le suivi des prêts sur toute la fenêtre d'observation.

2.1.1.2 Pertinence des variables

Pour chaque ligne de la base de données, les *reporting* fournissent un ensemble de 29 variables. Parmi celles-ci, toutes ne sont pas pertinentes à conserver pour une étude ultérieure. En effet, certaines ne seront pas exploitées du fait d'un nombre important de valeurs manquantes, les rendant inutiles pour une étude des taux de résiliation. Ainsi, dans le cadre de ce mémoire, les 12 variables conservées sont les suivantes :

- **Agence :** Nom anonymisé de l'agence du courtier ;
- **Nom :** Nom anonymisé de l'assuré ;
- **Prénom :** Prénom anonymisé de l'assuré ;
- **Date de Naissance :** Date de naissance de l'assuré ;
- **Date d'effet de garanties :** Date à partir de laquelle l'assuré est couvert par les garanties du contrat auquel il a souscrit ;
- **Couverture :** Types de garanties souscrites par l'assuré. Dans le cadre de ce contrat,il existe différentes formules couvrant le décès, l'incapacité et l'invalidité ;
- **Montant emprunté :** Montant emprunté en euros ;
- **Montant de cotisation annuel :** Coût annuel de l'assurance emprunteur ;
- **Durée :** Durée totale du prêt ;
- **Quotité :** Pourcentage du capital restant dû remboursé par l'assureur en cas de survenance du risque ;
- **Date de résiliation :** Date à laquelle le contrat a été résilié si résiliation il y a eu ;
- **Motif de résiliation :** Motif de la résiliation. Ces motifs sont multiples. Parmi eux nous avons notamment *remboursement total du prêt* ou encore *résiliation assuré.*

2.1.2 Nettoyage et retraitements effectués

Pour mener l'étude, il était nécessaire de réaliser des manipulations sur la base de données afin d'obtenir une base de travail saine. Nous avons donc nettoyé la base de ses défauts existants et appliqué un ensemble de retraitements afin d'obtenir des données cohérentes.

2.1.2.1 Création d'une base de données assurés

Dans un premier temps, il était nécessaire de créer une base de données annexe contenant des informations sur les assurés ayant souscrit un ou plusieurs prêts. Pour cela, il a fallu créer une clé primaire, que l'on nomme *Numéro assuré* , à partir des variables *Nom*, *Prénom* et *Date de Naissance* de la base de données de travail. Ensuite, certaines modifications ont été apportées afin de corriger des erreurs issues de la saisie des données en amont (décalage d'un jour au niveau de la date de naissance d'un assuré dans un *reporting* par exemple). Ainsi, nous obtenons une base de données annexe, contenant les variables *Nom*, *Prénom*, *Date de Naissance*, *Sexe*, et une variable *Numéro assuré* nouvellement créée et qui fait office de clé primaire pour cette base.

2.1.2.2 Création d'une base affaire

Nous entendons ici par affaire la souscription de plusieurs prêts à un même moment par une seule personne. Par exemple, celle-ci peut emprunter la somme de 300 000 € en trois prêts de 100 000 €. Ce choix de contracter plusieurs prêts au lieu d'un seul peut résulter de taux avantageux proposés par la banque jusqu'à un certain plafond ou une certaine durée comme les prêts réglementés à taux zéro. En pratique, les caractéristiques des prêts au sein d'une même affaire sont tellement différentes qu'il est impossible de les regrouper en un seul prêt commun représentatif de l'affaire qui serait exploitable pour une étude. Néanmoins, ce regroupement permet d'obtenir des informations intéressantes, d'un point de vue statistiques, sur la composition de notre portefeuille d'assurés et notamment sur les montants empruntés ou encore les âges à la souscription.

Pour obtenir cette base affaire, nous créons une clé primaire à partir de la variable *Date d'effet garantie*, déjà présente dans la base de données initiale, et de la variable *Numéro assuré* nouvellement créée.

2.1.2.3 Création de la base des prêts

Pour cette base, nous réunissons les informations permettant d'identifier de manière unique les informations de chaque prêt. Les courtiers saisissant les variables de manières différentes, il a été nécessaire d'utiliser une méthode générique afin de pouvoir rassembler tous les prêts. Ainsi, pour créer la base des prêts, nous récupérons tous les prêts d'une affaire uniquement à la dernière date d'observation. C'est pourquoi nous sommes en mesure de récupérer à la fois les informations concernant le prêt en lui même (durée, montant emprunté, garanties souscrites, ...) mais aussi les informations concernant son état (motif de résiliation, date de résiliation).

2.1.2.4 Suppression des prêts non pertinents

Une fois la base des prêts obtenue, il faut supprimer ceux qui n'ont finalement pas été souscrits par les assurés. Les motifs des prêts supprimés sont les suivants :

- **Adhésion résiliée pour tous les prêts :** l'assuré annule son adhésion sur tous les prêts d'une affaire avant le lancement effectif de celle-ci ;
- **Adhésion sans objet :** le motif d'adhésion n'étant pas spécifié le prêt n'est pas assuré ;
- **Déclaration inexacte à l'adhésion :** les informations transmises par l'assuré se sont avérées inexactes entraînant une annulation de la demande de prêt ;
- **Modification des garanties :** modification du niveau de couverture choisi avant la signature définitive du contrat. Ce choix entraîne généralement l'ouverture d'un nouveau contrat d'assurance avec les nouvelles garanties ;
- **Offre de prêt sans suite :** une offre a été faite mais l'assuré n'a pas souhaité y souscrire ;
- **Prêt non assurable :** l'assureur n'a pas pu transmettre une offre valide pour les caractéristiques de ce prêt ;
- **Prêt sans objet ouvert par erreur :** création d'une ligne de prêt résultant d'une erreur de saisie du courtier ;
- **Refus banque :** offre de prêt refusée par la banque ;
- **Renonciation assuré :** contrat non souscrit par l'assuré ;
- **Sans suite demande de pièces :** annulation du contrat suite à un manque de pièces justificatives dans le dossier.

2.1.2.5 Ajout de nouvelles variables d'études

Pour la création des trois bases mentionnées précédemment ainsi que pour l'étude qui va être réalisée, des variables ont dû être créées afin de rassembler de manière explicite des informations essentielles. Ainsi, nous sommes amenés à considérer les variables supplémentaires suivantes :

- **Ancienneté :** correspond au temps passé entre la date d'observation et la date de début d'effet des garanties. Nous l'obtenons de la manière suivante :

$$\textit{Ancienneté} = \textit{Date d'observation - Date d'effet de garanties}$$

Si le contrat est résilié avant la première date d'observation on fixe l'ancienneté à 0 ;

- **Date minimale d'observation :** correspond à la première date à laquelle une affaire est observée dans la base de données ;
- **Date maximale d'observation :** correspond à la dernière date à laquelle une affaire est observée dans la base de données. Même si au sein d'une affaire un prêt est terminé depuis longtemps il sera toujours présent dans la base de données à cette date ;
- **Censure** [1] **:** variable binaire déterminant si la donnée est censurée ou non. Nous affecterons la valeur 0 aux données censurées et la valeur 1 aux données non censurées. La définition de la censure sera donnée plus tard dans ce mémoire ;
- **Ancienneté au moment de la résiliation :** en cas de résiliation d'un prêt cette variable correspond au temps passé entre la date d'effet des garanties et la date de résiliation. Ainsi, nous obtenons l'ancienneté au moment de la résiliation de la manière suivante :

$$\textit{Ancienneté résiliation} = \textit{Date de résiliation - Date d'effet des garanties}$$

- **Montant total de l'affaire :** correspond à la somme des montants empruntés pour les prêts rattachés à une même affaire.

2.2 Statistiques sur la base de données

Une fois les bases des assurés, des affaires et des prêts obtenues, il peut être intéressant d'en extraire quelques statistiques afin d'appréhender au mieux le portefeuille que nous avons à notre disposition. Ainsi, cette section aura pour objectif d'essayer de décrire les particularités des données que nous allons étudier par la suite.

2.2.1 Statistiques sur la base des assurés

Concernant la base des assurés, les statistiques que nous pouvons en tirer sont assez sommaires. En effet, nous pouvons constater que nous avons à notre disposition un portefeuille de 2000 assurés. Au sein de ce portefeuille, les hommes et les femmes suivent la répartition suivante :

1. cf. partie 2.3.1 pour la définition

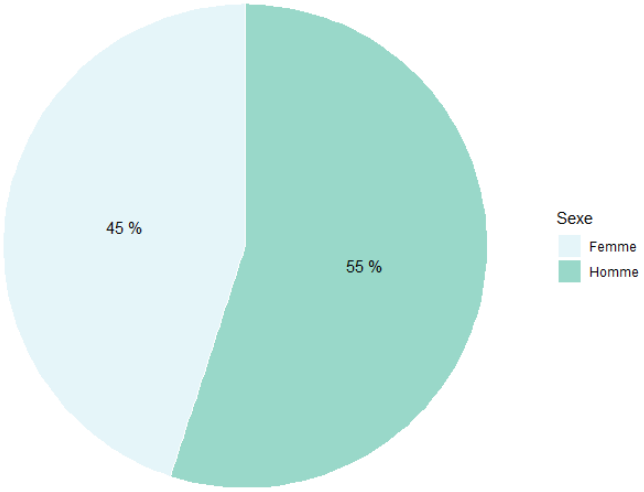


FIGURE 2.1 – Répartition par sexe dans le portefeuille

Nous observons une légère disparité dans le portefeuille avec une part plus importante d'hommes à hauteur de 55 % soit 1100 hommes pour 900 femmes.

2.2.2 Statistiques sur la base des affaires

En tant que telle, la base des affaires n'apporte pas d'informations importantes pour le calcul des taux de résiliation. Néanmoins, une étude de ses caractéristiques permet d'en apprendre davantage sur la composition de notre portefeuille et sur le comportement des assurés. Contrairement à la base des prêts, nous disposons ici d'une vision globale affaire par affaire. Ce point de vue permet d'avoir une appréciation claire du profil d'investissement des assurés.

Nous avons donc à disposition une base des affaires qui regroupe environ 2400 affaires pour un montant moyen emprunté par affaire de 260 000 €. Cette somme est légèrement supérieure aux montants qu'il est possible d'observer sur le marché de l'emprunt. Nous pouvons également noter 1,50 prêts par affaire en moyenne qui se répartissent de la manière suivante :

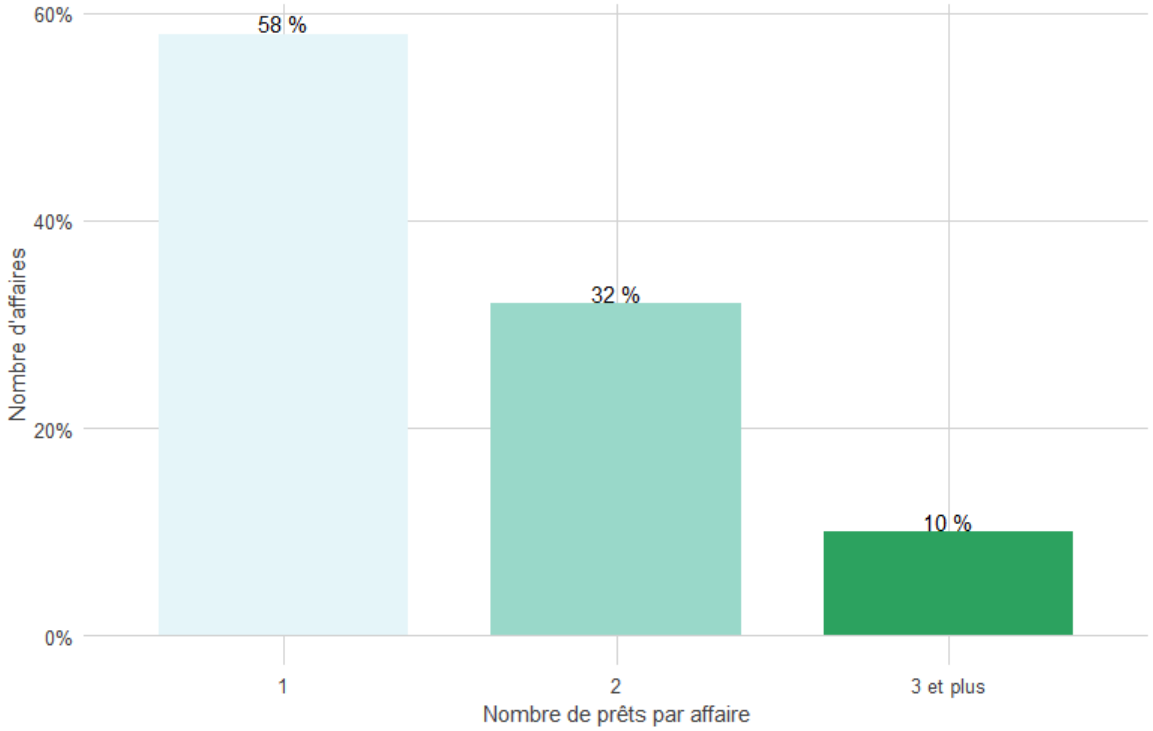


FIGURE 2.2 – Répartition du nombre de prêts par affaire

Nous pouvons constater que la plupart des affaires comprennent au maximum 2 prêts. Pour les affaires ayant 3 prêts ou plus, ce sont des prêts à faibles montants et faibles durées qui viennent généralement compléter un prêt plus important.

2.2.3 Statistiques sur la base des prêts

Concernant la base des prêts, nous avons à notre disposition un ensemble de 3400 prêts dont le capital initial sous risque moyen par prêt est de 120 000 € et dont les garanties se répartissent de la manière suivante :

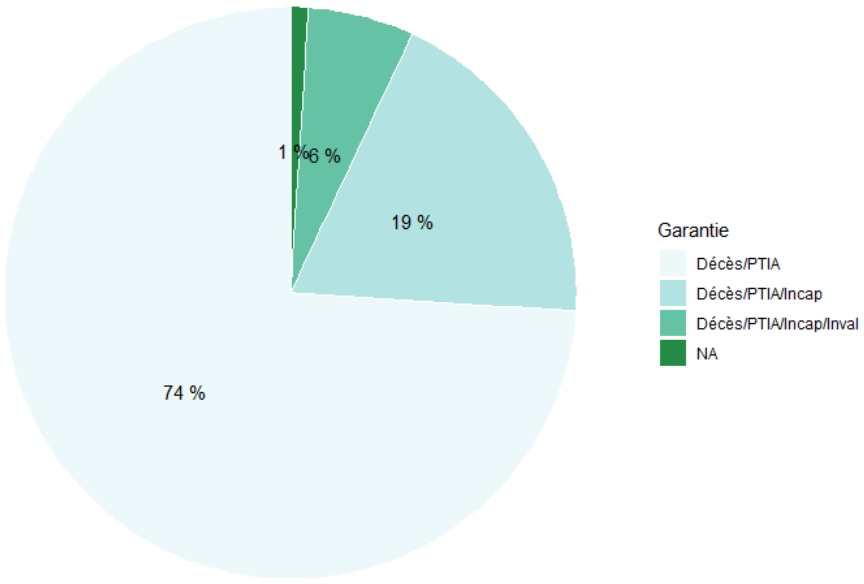


FIGURE 2.3 – Répartition des garanties

L'âge moyen à la souscription est aussi une donnée importante à prendre en compte. Pour notre portefeuille il est de 34 ans pour une durée moyenne d'emprunt de 16 ans. Nous avons donc à disposition un portefeuille relativement jeune par rapport au marché et qui semble être principalement constitué de prêts immobiliers, compte tenu des durées d'emprunt élevées. C'est à partir de cette variable d'âge à la souscription que nous avons décidé de faire une étude complémentaire de la résiliation en segmentant notre portefeuille en 3 classes d'âge qui ont les caractéristiques suivantes :

Classe d'âge :	**[20 ; 30 [**	**[30 ; 40 [**	**[40 ; 75 [**
Nombre de prêts :	800 prêts	2000 prêts	600 prêts
Âge moyen à la souscription :	27 ans	34 ans	45 ans
Durée moyenne d'emprunt :	18 ans	16 ans	14 ans

TABLE 2.1 – Tableau récapitulatif de la segmentation retenue

Le choix de cette segmentation provient de l'hypothèse selon laquelle les assurés peuvent être placés dans trois classes d'âge différentes. En supposant qu'une classe d'âge est représentative d'une génération à partir de dix âges consécutifs différents, la segmentation retenue est alors celle présentant la répartition la plus homogène.

2.3 Estimateur de Kaplan Meier

2.3.1 Notion de censure

En assurance emprunteur un prêt s'étend sur une grande période temporelle pouvant atteindre une trentaine d'années. Cette contrainte rend donc complexe l'observation d'un prêt sur toute sa période de vie. La censure des données devient alors un problème important à prendre en compte dans toute étude relative à la durée de vie des prêts. Dans le cadre de notre étude nous serons donc amenés à prendre en compte le phénomène de censure à droite.

Dans le cas de l'utilisation de l'estimateur de Kaplan-Meier, considérons une fenêtre d'observation $[\![A; B]\!]$ délimitant les dates d'observation contenues dans la base de données. Nous pouvons alors distinguer deux causes de censure à droite qui sont les suivantes :

- Le prêt est toujours en cours à la date B. Dans ce cas, le prêt n'a pas été résilié pendant la période d'observation et sa cause de sortie s'est produite sur une période non observable et donc inconnue ;
- Le prêt se termine pour une cause autre que celle étudiée, ici la résiliation du contrat. Les causes de sortie autre que le rachat peuvent être diverses : décès de l'assuré, prêt parvenu à son terme... ;

Nous pouvons résumer les différents types de données que nous trouvons dans la base considérée selon les schémas suivants :

- Donnée complète :

- Donnée censurée à droite (hors de la fenêtre d'observation) :

– Donnée censurée à droite (dans la fenêtre d'observation) :

Pour ces trois schémas nous avons la légende suivante :

– A : date de début d'observation ;

– B : date de fin d'observation ;

– S : sortie pour la cause étudiée ;

– X : sortie pour une cause autre que celle étudiée.

La notion de troncature à gauche est également généralement associé à l'estimateur de Kaplan Meier. Une donnée X est dite tronquée si X < A ou bien si X > B. Dans le cadre de cette étude, les données tronquées correspondent aux données qui n'appartiennent pas à la fenêtre d'observation et donc à la base de données. Par la suite, ces données ne seront pas prises en compte dans le calcul des estimateurs de Kaplan-Meier.

2.3.2 Estimateur et intervalles de confiance bootstrap

Pour cette étude, le choix a été fait de calculer des taux annuels de résiliation selon l'approche de Kaplan-Meier [Kaplan et Meier, 1958]. Une approche mensuelle, certes plus fine, aurait pu être envisagée mais les données ne permettent pas d'obtenir un taux de résiliation pour chaque mois d'ancienneté.

Il existe de nombreuses formules, toutes équivalentes, de l'estimateur de Kaplan-Meier. Pour notre étude, nous conserverons la formulation suivante :

$$\forall t \geq 0, \widehat{S}_{KM}(t) = \prod_{i, X_{(i)} \leq t}^{n} \left(\frac{n-i}{n-i+1} \right)^{\delta_{(i)}}.$$

Où :

– X_i est l'ancienneté à laquelle le prêt i quitte l'échantillon d'étude quelle que soit la cause ;

- $X_{(1)}, ..., X_{(n)}$ est l'échantillon ordonné des anciennetés auxquelles les prêts quittent l'échantillon quelle que soit la cause ;
- $\delta_{(i)}$ la variable binaire valant 0 si $X_{(i)}$ est censurée et 1 sinon ;
- n la taille de l'échantillon de données de départ ;

La taille de la base de données à notre disposition étant relativement modeste, l'utilisation des formules fermées des intervalles de confiance ne permet pas de respecter les conditions minimales de robustesse des tests statistiques qui y sont associés. Par conséquent, une méthode de calcul de cet estimateur par bootstrap a été employée.

La méthode bootstrap consiste à appliquer la formule de l'estimateur de Kaplan-Meier précédemment mentionnée à un grand nombre d'échantillons de données. Ces échantillons, de la même taille que l'échantillon de départ, sont obtenus par un tirage aléatoire avec remise sur l'échantillon de départ. Afin d'obtenir les estimations des taux de survie ainsi que les intervalles de confiance asymptotiques associés suffisamment robustes, il est nécessaire de prendre un nombre d'échantillons bootstrap important. Pour notre étude, ce nombre est fixé à N = 1000.

Ainsi, pour finalement obtenir les informations nécessaires pour la suite de l'étude, nous procédons de la manière suivante :

- Pour obtenir les estimations des taux de survie de Kaplan-Meier, nous prenons la moyenne de tous les taux calculés dans l'échantillon obtenu par la méthode bootstrap ;
- Pour obtenir les bornes des intervalles de confiance asymptotiques à 1-α, nous prenons respectivement les quantiles $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de l'échantillon obtenu par la méthode bootstrap. Dans notre étude, nous prendrons $\alpha = 5\%$

Ces intervalles de confiance permettront par la suite de déterminer la validité des lissages de Whittaker-Henderson qui seront réalisés.

2.3.3 Résultats obtenus

Nous avons donc appliqué l'estimateur de Kaplan-Meier aux données non segmentées. Ainsi, nous obtenons le graphique suivant représentant les estimations des taux bruts de survie :

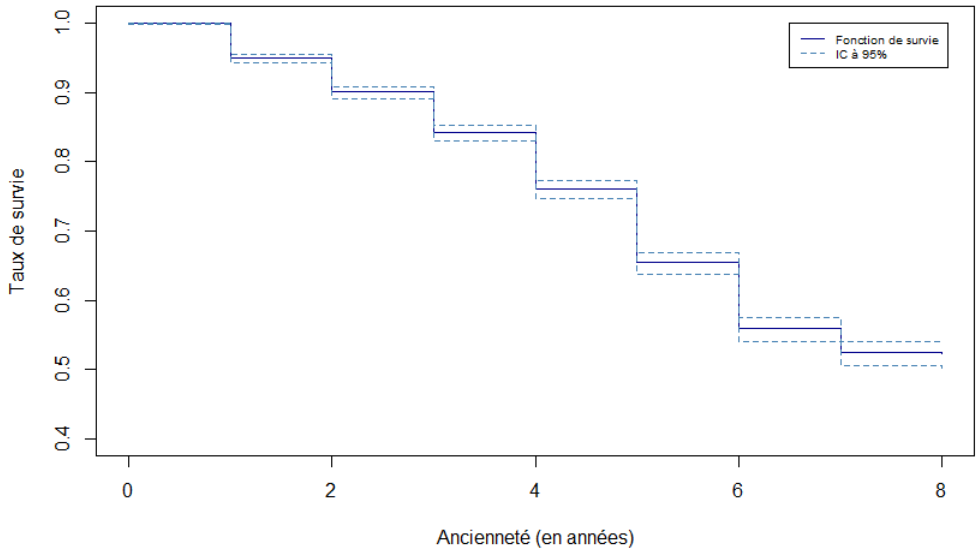


FIGURE 2.4 – Estimateur de Kaplan-Meier sans segmentation

Ici, nous pouvons voir que les intervalles de confiance autour des estimations des taux bruts de survie s'élargissent à mesure que l'ancienneté augmente. Cette augmentation s'explique par une exposition qui diminue au fil du temps. Néanmoins, même pour l'ancienneté la plus grande, l'intervalle de confiance reste relativement restreint. De ce fait, les taux de résiliation ne peuvent prendre qu'un nombre limité de valeurs.

L'estimateur de Kaplan-Meier a également été calculé sur les trois classes retenues pour la segmentation. Dans un souci de clarté, seuls les résultats obtenus pour la classe des 20-30 ans seront exposés :

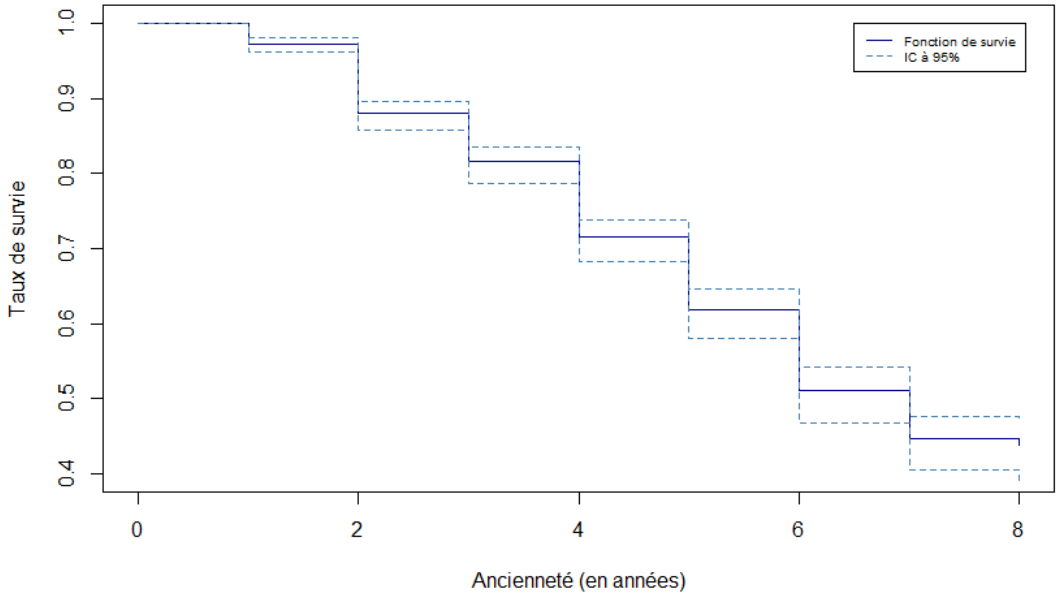


FIGURE 2.5 – Estimateur de Kaplan-Meier pour la classe 20-30 ans

Là aussi, le même phénomène d'agrandissement des intervalles de confiance est observé. Toutefois, ces écarts sont beaucoup plus importants que pour l'étude sans segmentation. En effet, l'étude segment par segment fait diminuer l'exposition et par conséquent fait augmenter l'incertitude autour des estimations des taux de survie. Il faudra donc apporter une attention particulière aux lissages réalisés dans la suite de l'étude, notamment pour les segments où l'exposition est la plus faible.

Nous allons désormais appliquer une méthode de lissage aux estimations des taux bruts de survie que nous venons d'obtenir. Pour cela, nous avons fait le choix d'appliquer le lissage non paramétrique de Whittaker-Henderson.

2.4 Lissage de Whittaker-Henderson

2.4.1 Principe

Le lissage de Whittaker-Henderson [Lee Giesecke, 1981] appartient à la famille des lissages non paramétriques. Il repose sur la combinaison linéaire d'un critère de régularité et d'un critère de fidélité. L'objectif est de trouver des valeurs estimées qui minimisent cette combinaison linéaire. L'essentiel de la complexité de ce lissage est la dualité qui existe entre la fidélité et la régularité. Une diminution du critère de fidélité entraîne une augmentation du critère de régularité et inversement.

2.4.1.1 Critère de fidélité

Considérons une courbe $c : x \to c_x$. La courbe c est d'autant plus fidèle aux taux bruts $\widehat{q}_x$ que le critère $F(c)$ est faible. $F(c)$ est défini de la manière suivante :

$$F(c) = \sum_{x=1}^{x_{max}} w_x (c_x - \widehat{q}_x)^2.$$

Où :

- $\widehat{q}_{x_1}, ..., \widehat{q}_{x_{max}}$ sont les taux bruts estimés par la méthode de Kaplan-Meier ;
- $c_1, ..., c_{max}$ sont les taux lissés estimés ;
- $w_1, ..., w_{max}$ sont des poids positifs ;

Le choix des poids $w_1, ..., w_{max}$ dépend de l'importance que nous voulons donner à chaque taux. Un choix classique revient à donner le même poids pour chaque taux mais nous pourrions considérer des poids plus importants sur les taux qui ont une exposition plus grande que les autres. Dans le cadre de notre étude, nous avons fait le choix de considérer un poids identique pour tous les taux.

2.4.1.2 Critère de régularité

Nous prenons en compte la même courbe c définie dans la partie sur le critère de fidélité. La courbe c est d'autant plus régulière que le critère $S(c)$ est faible. $S(c)$ est défini de la manière suivante :

$$S(c) = \sum_{x=1}^{x_{max}-z} [(\Delta^z c)_x]^2.$$

Où :

- Δ est l'opérateur défini par $(\Delta c)_x = c_{x+1} - c_x$ pour $x \in \{1, ..., x_{max} - 1\}$;
- z est un paramètre d'ordre. Usuellement, ce paramètre vaut 1 ou 2 ;
- $\Delta^z = \Delta(\Delta^{z-1})$;

2.4.1.3 Critère de Whittaker-Henderson

Ainsi, avec ces deux critères, nous pouvons définir le critère de Whittaker-Henderson $WH_h(c)$ de la manière suivante :

$$WH_h(c) = F(c) + h * S(c).$$

L'objectif principal du lissage de Whittaker-Henderson est de minimiser ce critère en fonction de la valeur du paramètre h. La valeur de ce paramètre permet de donner plus ou moins d'importance au critère de régularité. Plus la valeur de h est élevée, plus les taux lissés seront réguliers. Un h fixé à 0 correspond au cas où il n'y a pas de différence entre les taux bruts et les taux lissés.

2.4.2 Application aux estimations des taux bruts de survie

Dans le cadre de notre étude, nous allons appliquer le lissage de Whittaker-Henderson aux estimations des taux bruts calculés avec l'estimateur de Kaplan-Meier dans la partie précédente. Afin d'avoir plusieurs approches, nous testerons différents paramétrages obtenus après un premier examen graphique sur toutes les segmentations réalisées. Dans un souci de lisibilité, seuls les lissages réalisés sur l'étude sans segmentation ainsi que sur la classe d'âge des 20-30 ans seront présentés. Ainsi, nous obtenons les lissages suivants :

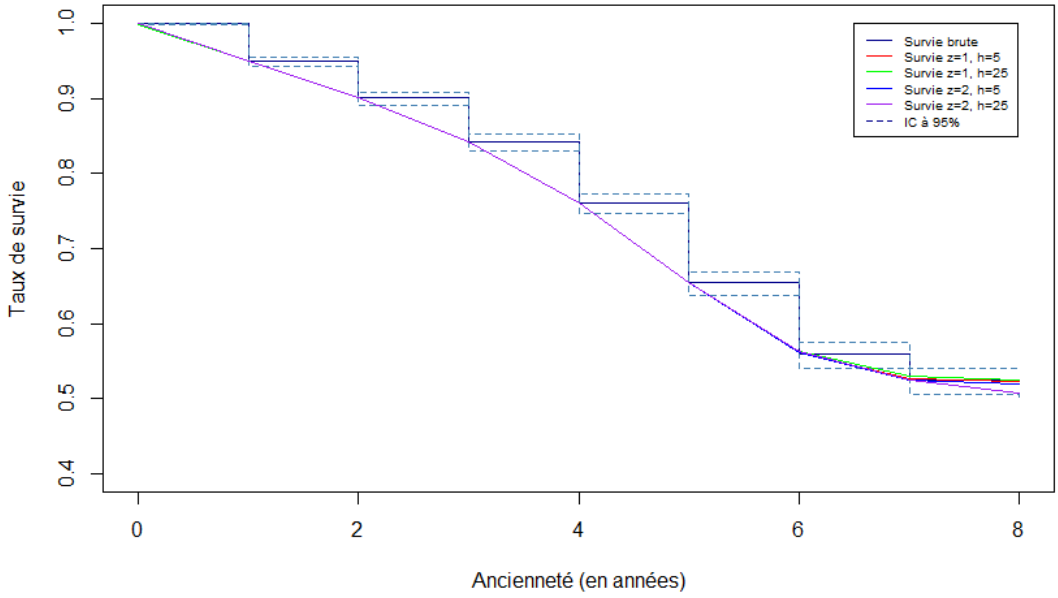


FIGURE 2.6 – Lissages de Whittaker-Henderson sans segmentation

Pour que les paramètres testés soient retenus, il est nécessaire que tous les points calculés appartiennent aux intervalles de confiance obtenus avec la méthode *bootstrap*. Les résultats obtenus peuvent être résumés dans le tableau suivant :

Paramètres	Sans segmentation	Classe d'âge 20-30 ans
z=1, h=5	Oui	Oui
z=1, h=25	Oui	Oui
z=2, h=5	Oui	Oui
z=2, h=25	Oui	Non

TABLE 2.2 – Appartenance des taux lissés de survie aux intervalles de confiance

En l'état actuel des choses, il est impossible de choisir un unique lissage. Le meilleur lissage sera donc obtenu par un passage aux taux de résiliation.

Pour obtenir les taux de résiliation à partir des taux lissés de survie, la formule suivante est utilisée :

$$\forall x \in \{1, ..., x_{max} - 1\}, \widehat{r}(x) = 1 - \frac{\widehat{q_l}(x+1)}{\widehat{q_l}(x)},$$

Où :

– $\widehat{r}(x)$ le taux de résiliation à l'ancienneté x ;

– $\widehat{q_l}(x)$ le taux de survie lissé à l'ancienneté x ;

Pour choisir le meilleur lissage, une méthode prudente a été utilisée. En effet, la résiliation d'un contrat d'assurance emprunteur mettant fin aux versements futurs des primes, plus un taux de résiliation est élevé, plus la rentabilité du portefeuille risque d'être impactée négativement. Les taux observés sur les trois anciennetés les plus grandes étant déterminant pour l'étape de prolongement des taux, nous avons fait le choix de conserver les paramètres du lissage présentant les taux de résiliation les plus élevés sur ces trois anciennetés. Ainsi, pour les deux segmentations considérées, nous obtenons les paramètres suivants :

Segmentation	Paramètre z	Paramètre h
Sans segmentation	2	25
Classe 20-30 ans	2	5

TABLE 2.3 – Paramètres des lissages retenus

Une fois les paramètres du lissage sélectionnés, la dernière étape consiste à prolonger les taux jusqu'à 30 ans afin d'obtenir une loi de résiliation complète. Cette étape fera l'objet de la section suivante.

2.5 Fermeture des taux de résiliation

2.5.1 Problématiques rencontrées et besoins

2.5.1.1 Allure des taux

D'un point de vue pratique, l'allure de la courbe des taux de résiliation est assez intuitive. Nous pouvons observer une première phase où les taux de résiliation sont croissants puis, une seconde phase de décroissance jusqu'au terme du prêt. Ce comportement s'explique facilement par l'intérêt que peut avoir un assuré à résilier son prêt. L'âge étant la variable ayant le plus d'influence sur le prix du contrat emprunteur, il est généralement plus intéressant pour l'assuré de chercher une offre d'assurance plus adaptée à son profil de risque au début du prêt. A mesure que le temps passe, le profil de l'assuré devient plus risqué et les capitaux assurés plus faibles. Par conséquent, il aurait moins d'intérêt économique à résilier son contrat.

Dans le cadre de notre étude, les taux de résiliation que nous avons calculés sont représentatifs du phénomène décrit ci-dessus. Néanmoins, les données extraites de la base de données ne permettent pas de calculer des taux de résiliation au delà de 8 ans. Il nous faut donc trouver une fonction qui permette de fermer les taux obtenus en respectant les contraintes de comportement des assurés.

2.5.1.2 Contrôle de la prudence

Dans le cadre de la construction des taux de résiliation, il est important de pouvoir choisir avec quel degré de prudence nous souhaitons poursuivre l'étude. Cela implique que la méthode avec laquelle nous allons réaliser la fermeture des taux doit intégrer un paramètre permettant de contrôler cet aspect.

2.5.2 Solutions apportées

Pour fermer les taux de résiliation précédemment calculés, nous avons fait le choix d'utiliser la fonction de survie de Weibull. Cette fonction est de la forme suivante :

$$\forall t > 0, S(t) = e^{-\lambda \cdot t^{\alpha}}.$$

Où :

- λ est un paramètre défini à partir du point de raccordement ;
- α est un paramètre permettant d'ajuster la forme de la courbe pour obtenir un raccordement cohérent et contrôler la prudence ;

Cette fonction présente l'avantage de répondre à toutes les problématiques auxquelles nous faisons face. La forme même de la fonction répond aux attentes de décroissance à mesure que l'ancienneté augmente tout en laissant la possibilité d'ajuster les taux pour avoir une approche prudente.
Ainsi pour déterminer la valeur du paramètre λ nous obtenons, au point de raccordement, les équivalences suivantes :

$$\begin{aligned} \widehat{q}_n &= S(n) \\ &= e^{-\lambda \cdot n^{\alpha}} \end{aligned}$$

Par conséquent, nous avons :

$$\lambda = -\frac{ln(\widehat{q}_n)}{n^{\alpha}}$$

La dernière étape consistera à tester différentes valeurs du paramètre α. Plus α est grand, plus les taux vont converger rapidement vers 0. Nous chercherons donc à sélectionner la valeur du paramètre α qui permet d'obtenir la fermeture la plus lisse possible au niveau du point de raccordement, tout en gardant un niveau de prudence satisfaisant.

2.5.3 Choix et mise en commun

Le graphique suivant illustre les taux de résiliation que nous obtenons pour l'étude sans segmentation :

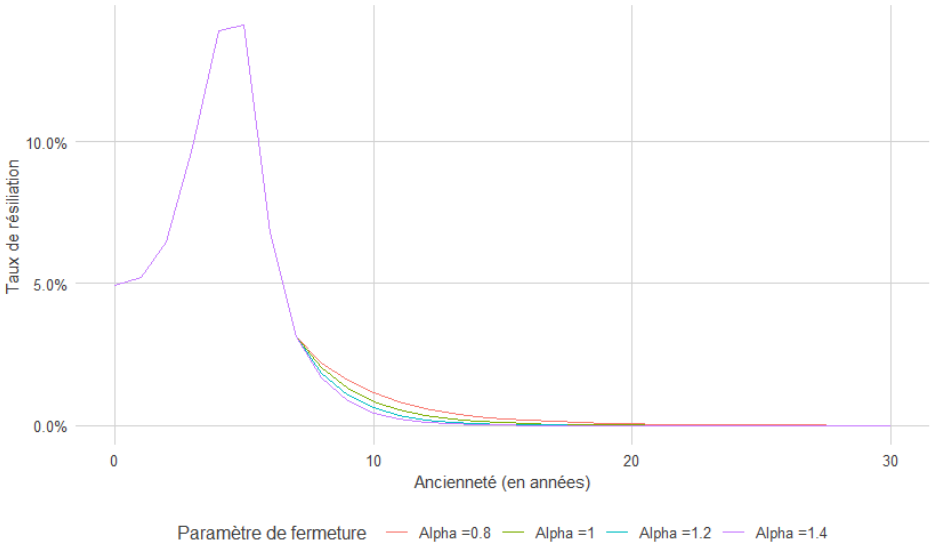


FIGURE 2.7 – Fermetures des taux de résiliation sans segmentation

Pour cette étude, nous avons fait le choix de conserver la fermeture avec le paramètre $\alpha = 1$. En effet, plusieurs paramètres semblaient donner une fermeture lisse au niveau du point de raccordement. Cependant, c'est ce choix de paramètre qui présente les taux de résiliation les plus élevés pour les anciennetés les plus élevées. La courbe associée au paramètre $\alpha = 0.8$ ne correspondant pas au critère de cohérence au point de raccordement, c'est donc le paramétrage donnant la plus grande prudence.

Pour la classe d'âge 20-30 ans, nous avons également les différents paramétrages suivants :

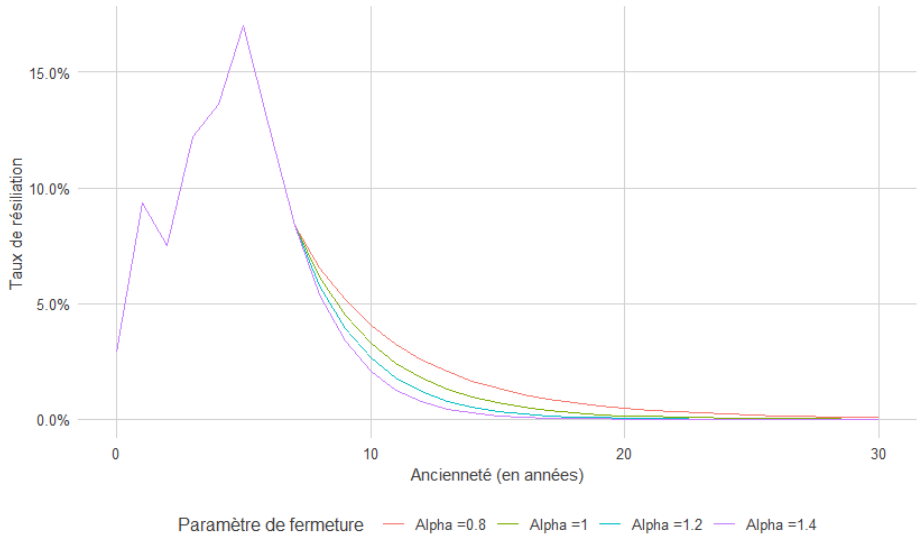


FIGURE 2.8 – Fermeture des taux de résiliation pour la classe d'âge 20-30 ans

Ici, nous avons conservé le paramètre $\alpha = 1$ pour les mêmes raisons que pour l'étude sans segmentation. De manière générale, la méthode appliquée aux autres classes d'âge est la même et les paramètres α retenus sont 1.

Par conséquent, en mettant toutes les courbes des taux de résiliation sur le même graphique, nous obtenons les résultats suivants :

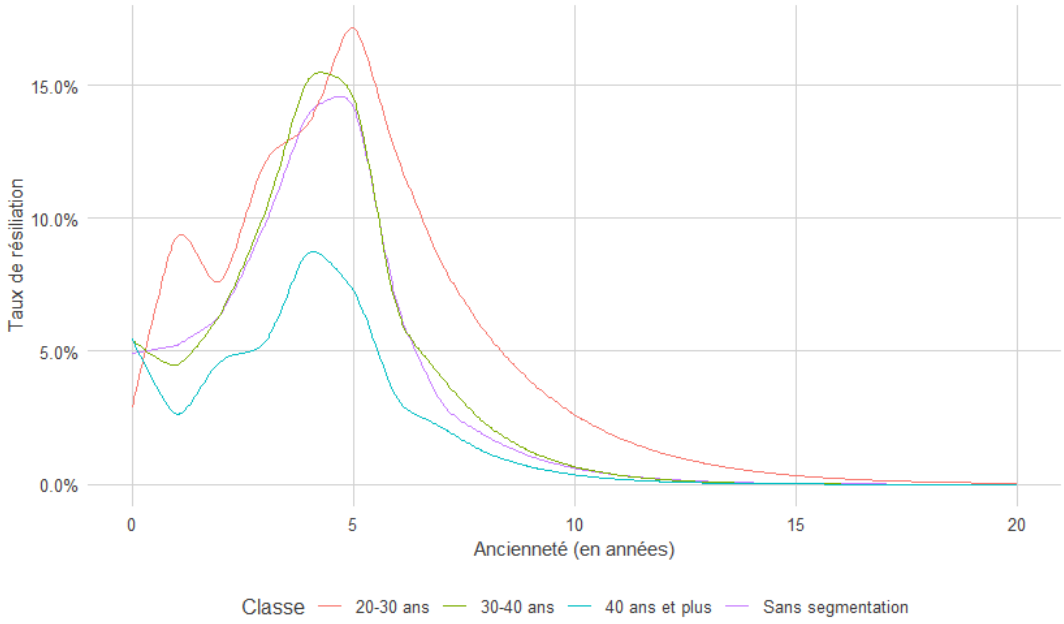


FIGURE 2.9 – Taux de résiliation retenus pour toutes les segmentations

Ces résultats mettent en valeur les différents comportements auxquels nous pouvions nous attendre à obtenir. En effet, la classe d'âge des plus de 40 ans contenant les profils les plus âgés est plus exposée au risque décès, les personnes appartenant à cette classe vont se voir proposer des tarifs d'assurance plus élevés que les profils les plus jeunes, s'ils cherchent à en changer. Par conséquent, ils seront sont moins enclins à résilier leur contrat. A l'inverse, les profils les plus jeunes, appartenant à la classe d'âge des 20-30 ans, se verront proposer des tarifs plus avantageux et seront donc plus disposés à changer d'assurance. De plus, nous pouvons observer une cohérence dans les résultats obtenus : les taux de résiliation de l'étude sans segmentation sont une "moyenne" des taux des trois autres classes d'âge réunies.

2.5.4 Interprétation des variations

Les courbes des taux de résiliation obtenues à l'aide des méthodes de Kaplan-Meier et de Whittaker-Henderson présentent des irrégularités notamment au niveau des premières an-

ciennetés. Là où des taux croissants seraient attendus, il est possible d'observer des croissances et décroissances successives. Ces variations pourraient s'expliquer par divers éléments.

La baisse des taux d'emprunt consécutive à la crise de 2008 aurait entraîné une hausse du rachat conjoncturel. Le rachat conjoncturel correspond à un rachat lié à l'évolution de l'environnement économique. Le pendant du rachat conjoncturel est le rachat structurel. Ce type de rachat correspond à un rachat lié aux caractéristiques propres de l'emprunt. L'hypothèse d'une hausse du rachat conjoncturel est renforcée par une sinistralité importante au niveau de l'année 2015, du point de vue de la résiliation, quelle que soit l'année de souscription comme le montre le tableau suivant :

	Ancienneté							
Année souscription	0	1	2	3	4	5	6	7
2009	3,2%	10,6%	4,6%	8,1%	7,8%	15,6%	4,2%	5,2%
2010	5,3%	1,8%	2,6%	6,4%	17,1%	18%	6,7%	
2011	4%	2,7%	8,2%	8,7%	12,9%	9,8%		
2012	6,6%	7,6%	13,6%	16,2%	9,9%			
2013	3,2%	6,3%	6,3%	17,9%				
2014	4,9%	14,1%	2,38%					
2015	6,6%	3,6%						
2016	0%							

TABLE 2.4 – Résiliation observée par ancienneté et par année de souscription

La majorité des prêts de notre base ayant été souscrits entre 2011 et 2013, la sur-sinistralité se retrouve donc sur les anciennetés les plus faibles. Les variations relevées pour les anciennetés les plus faibles pourraient également s'expliquer par la promulgation de la loi Hamon permettant la résiliation de son contrat emprunteur à tout moment avant la première date d'anniversaire. La taille du jeu de données et les faibles expositions qui en découlent peuvent également être des facteurs explicatifs de ces irrégularités.

Un lissage beaucoup plus important aurait pu être envisagé notamment sur les anciennetés les plus faibles. Toutefois, l'objectif de cette étude étant de comparer une méthode classique avec des méthodes de *Machine Learning*, le choix retenu consiste à les conserver tels quels et de voir si ces nouvelles méthodes permettent de lisser ces irrégularités. L'objectif est également de ne pas biaiser les résultats obtenus par la méthode classique. Les méthodes de *Machine Learning* feront l'objet du chapitre suivant.

Chapitre 3

Modélisation Data Science

Ce chapitre a pour but de détailler les méthodes de *Machine Learning* employées afin de modéliser les taux de résiliation. Après avoir fait une présentation générale des concepts généraux de *Machine Learning*, les trois méthodes suivantes seront optimisées :

- Le modèle d'arbre de décision fondé sur la méthode CART (*Classification and And Regression Tree*) ;
- Le modèle de *Random Forest* fondé sur l'apprentissage d'un ensemble d'arbres de décision ;
- Le modèle XGBoost (*eXtreme Gradient Boosting*) fondé sur l'amélioration d'un arbre de décision par itérations successives.

Enfin, une fois les résultats obtenus, nous aboutirons à des considérations de robustesse nécessaire dans le cadre de la base de données employée ainsi qu'à une comparaison des modèles.

3.1 Mise en contexte

Dans cette partie, les éléments généraux relatifs à la modélisation des taux de résiliation via des méthodes de *Machine Learning* seront abordés. Après une présentation des méthodes employées, l'accent sera porté sur le langage utilisé ainsi que les retraitements apportés à la base de données.

3.1.1 Présentation générale

En *Machine Learning*, la pluralité des méthodes couvre différents aspects de la modélisation. Ces méthodes peuvent être classées en deux catégories :

- **L'apprentissage supervisé :** pour ces modèles, l'apprentissage est guidé par des étiquettes affectées au préalable aux données. Ces étiquettes sont les résultats qui vont chercher à être prédits à l'avenir. Le modèle a alors pour objectif de diminuer l'erreur entre le résultat prédit et le résultat attendu. Ces étiquettes peuvent être soit qualitatives soit quantitatives. Le terme de classification est utilisé pour les étiquettes qualitatives tandis que le terme de régression est utilisé pour les étiquettes quantitatives ;

- **L'apprentissage non supervisé :** pour ces modèles, l'apprentissage est totalement autonome. Les données sont entrées de manière brute, sans étiquette. L'objectif est alors de créer des groupes de données aux caractéristiques communes.

Dans le cadre de cette étude, les méthodes d'apprentissage supervisé et plus particulièrement les méthodes de classification seront étudiées. Il s'agira, ici, de prédire la variable binaire qui nous indique si le prêt est résilié ou non.

Les trois algorithmes qui seront confrontés sont les suivants :

- **Le modèle par arbre de décision :** ce modèle est fondé sur la méthode CART (*Classification And Regression Trees*) ;

- **Le modèle de *Random Forest* :** ce modèle est fondé sur l'apprentissage ensembliste d'arbres de décision ;

- **Le modèle XGBoost *(eXtreme Gradient Boosting)* :** ce modèle est fondé sur une amélioration du classificateur par itérations successives.

3.1.2 Langage et bibliothèques utilisés

Le langage de programmation Python dispose d'un grand nombre de bibliothèques spécialisées dans les méthodes de *Machine Learning*. Pour cette étude, les bibliothèques utilisées pour modéliser les taux de résiliation sont les suivantes :

- Pandas : cette librairie permet de faciliter la manipulation des bases de données ;

- Numpy : cette librairie permet de faciliter le calcul matriciel ;

- Graphviz : cette librairie permet d'améliorer la visualisation des arbres de décision ;

- Matplotlib et Seaborn : ces librairies permettent d'améliorer la visualisation des indicateurs de performance ;

- Scikit-learn : cette librairie permet d'implémenter les modèles d'arbre de décision et de *Random Forest* de même qu'elle intègre également les métriques de performance de ces modèles ;
- Xgboost : cette librairie permet d'implémenter le modèle *XGBoost* de même qu'elle intègre également les métriques de performance de ce modèle.

3.1.3 Modifications sur le jeu de données

Pour rappel, la base de données utilisée pour modéliser les taux de résiliation à l'aide des méthodes de *Machine Learning* est composée des variables explicatives suivantes :

- Agence ;
- Montant emprunté ;
- Montant affaire ;
- Année de naissance ;
- Année de souscription ;
- Nombre de prêts par affaire ;
- Quotité ;
- Couverture ;
- Montant cotisation annuelle ;
- Durée du prêt ;
- Périodicité de remboursement.

Le nombre et la qualité des variables de la base de données étant fortement limités, aucune étude complémentaire visant à sélectionner certaines variables au dépend d'autres n'a été réalisée. Par conséquent, l'ensemble des variables présentées ci-dessus vont servir à estimer la classe de la variable cible *Résiliation*.

Parmi les variables explicatives, deux types se distinguent :

- **Les variables quantitatives :** une variable est quantitative si elle est représentée par un nombre ;

– **Les variables qualitatives :** aussi appelées variables catégorielles, ces variables sont toutes celles qui ne sont pas quantitatives.

Afin que les méthodes de *Machine Learning* puissent exploiter les variables catégorielles, il est nécessaire de les retraiter au travers d'un codage disjonctif complet. Cette méthode permet de transformer les variables catégorielles en variables quantitatives.

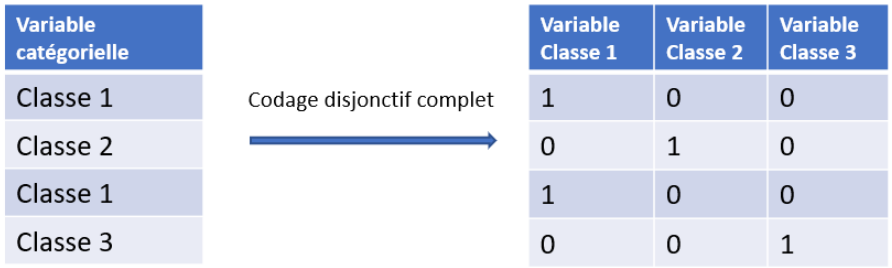


FIGURE 3.1 – Principe du codage disjonctif complet

Comme expliqué dans le schéma ci-dessus, le codage disjonctif complet consiste à créer autant de variables quantitatives binaires qu'il y a de classes dans la variable catégorielle initiale. Pour chaque instance, la valeur 1 est affectée à la variable quantitative représentant sa classe. La valeur 0 est affectée aux autres variables pour la même instance.

3.2 Arbre de décision

Dans cette partie, la modélisation des taux de résiliation à l'aide d'un arbre de décision va être abordée. Après une présentation des principes théoriques de cette méthode, les résultats obtenus seront explicités pour ensuite aboutir aux limites que pose cette méthode.

3.2.1 Principes théoriques

L'apprentissage par arbre de décision repose sur le modèle CART [Léo Breiman, 1996]. Ce modèle se base sur la classification des données par une suite de tests réalisés sur les variables explicatives. Ces tests sont effectués de manière hiérarchique et récursive. A chaque noeud de l'arbre, une variable est choisie afin de séparer l'échantillon initial en deux sous-échantillons. Les deux principaux critères de séparation des classes sont la variance intra-noeuds et le critère de Gini. En pratique, les deux critères donnent des résultats très similaires bien que les formules soient différentes. Dans la suite, nous utiliserons le critère de Gini.

En pratique, à chaque noeud, l'algorithme cherche à minimiser la quantité suivante :

$$J(k, t_k) = \frac{m_{gauche}}{m} G_{gauche} + \frac{m_{droite}}{m} G_{droite}$$

Où :

- k est la variable explicative considérée ;
- t_k est le seuil considéré ;
- $m_{gauche/droite}$ est le nombre d'instances dans le noeud gauche/droite ;
- m est le nombre d'instances dans l'échantillon initial ;
- $G_{gauche/droite}$ est l'indice de Gini du noeud gauche/droite.

Pour un noeud considéré, on obtient l'indice de Gini de la manière suivante :

$$G = 1 - \sum_k p_k^2$$

Où p_k est la proportion des instances de la classe k dans le noeud considéré.

L'indice de Gini vaut alors 0 dans le cas où un noeud ne contient que des instances d'une même classe. Quand chaque feuille ne contient plus qu'une seule instance, l'algorithme s'arrête : on parle alors d'arbre saturé.

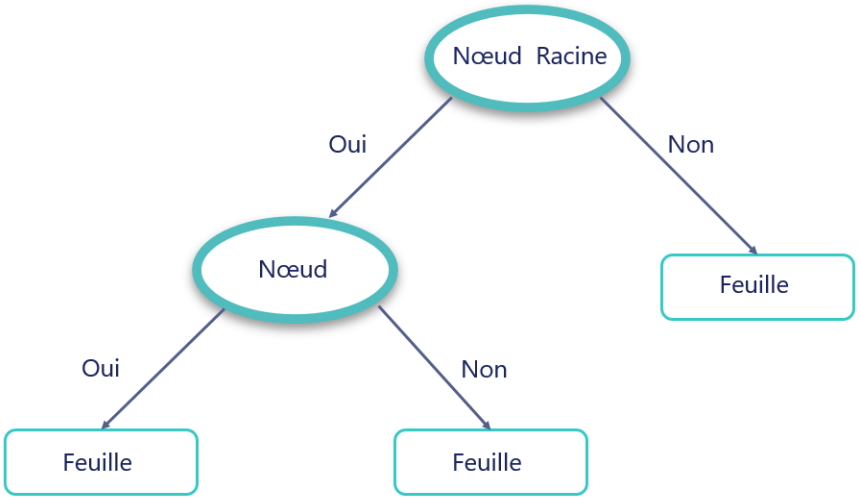


FIGURE 3.2 – Représentation graphique d'un arbre de décision

La figure ci-dessus représente le fonctionnement général d'un arbre. Ici, cet arbre a une profondeur de deux. Dès le premier noeud, l'algorithme arrive à créer une feuille avec toutes les instances appartenant à la même classe. Un second noeud est ensuite nécessaire pour séparer le jeu de données en deux feuilles distinctes.

3.2.2 Mise en pratique

Pour cette méthode, les différentes étapes permettant d'aboutir aux taux de résiliation seront abordées en détail. Pour les autres méthodes, celles-ci ne seront pas rappelées et seuls les résultats seront présentés.

3.2.2.1 Étape 1 : Séparation Train/Test

De manière générale, pour entraîner un modèle de *Machine Learning*, il est nécessaire de séparer la base de données en deux parties : d'une part la base d'entraînement et d'autre part la base de test. La séparation entre la base d'entraînement et la base de test est réalisée aléatoirement afin d'éviter les effets de structures liés à la construction de la base de données. Toutefois, pour cette étude, il est nécessaire d'ajouter une condition lors de cette séparation. En effet, étant donné que l'objectif est d'obtenir un taux pour chaque année d'ancienneté, les prêts doivent être équitablement répartis dans les deux bases. Un cas extrême, où les prêts les plus récents appartiendraient à la base d'entraînement et les prêts les plus anciens à la base de test, fausserait tous les résultats.

Le choix de la proportion de la base de données attribuée à la base d'entraînement et à la base de test est totalement arbitraire. Pour cette étude, la base d'entraînement représente 80% de la base de données initiale.

3.2.2.2 Étape 2 : Optimisation de la méthode

Avant l'entraînement des modèles de *Machine Learning*, il est nécessaire de fixer certains paramètres qui appliqueront des contraintes sur ces méthodes. Ces paramètres sont appelés des hyper-paramètres. Pour la plupart, les hyper-paramètres peuvent prendre une infinité de valeurs. Par conséquent, il n'est pas possible de tester toutes les combinaisons possibles. Une procédure d'exploration simple et exhaustive, nommée *Grid Search*, a été mise en place. Elle consiste à tester chaque modèle en faisant varier les hyper-paramètres sur des intervalles choisis préalablement formant alors une grille. Les hyper-paramètres optimaux sont ensuite sélectionnés à l'aide de cette méthode.

Pour sélectionner les hyper-paramètres optimaux, la méthode *Grid Search* fait appel au concept de cross-validation. Ce concept est également connu sous le nom de méthode *K-fold* où un *fold* désigne une sous partie de la base de données complète. Cette méthode permet de s'affranchir d'une base de validation. En effet, c'est sur la base d'entraînement que les modèles vont être entraînés et validés pour ensuite être comparés.

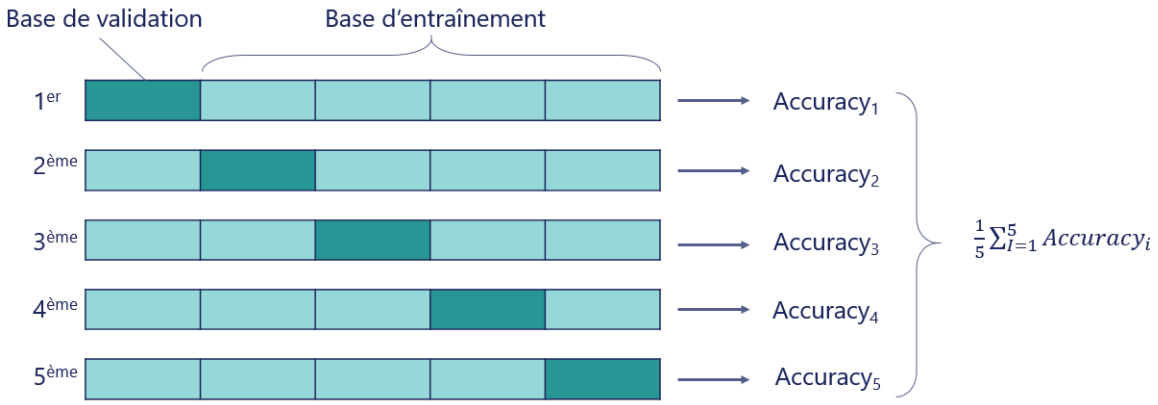


FIGURE 3.3 – Principe de la cross validation

Considérons une cross-validation à cinq *fold*. Comme expliqué dans le schéma ci-dessus, la base d'entraînement est séparée en 5 blocs distincts. Les modèles sont ensuite entraînés sur quatre des cinq blocs puis testés sur le bloc restant. L'opération est répétée de manière à ce que chaque bloc serve une fois de base de test. Les indicateurs de performance sont ensuite moyennés. La meilleure combinaison d'hyper-paramètres est alors le modèle qui obtient le meilleur score moyen par cross-validation.
Pour la modélisation par arbre de décision, trois hyper-paramètres seront optimisés. Ces hyper-paramètres sont les suivants :

- *max_depth* : cet hyper-paramètre correspond à la profondeur maximale que peut avoir l'arbre ;

– *min_samples_split* : cet hyper-paramètre correspond au nombre d'instances minimum qu'un noeud doit avoir pour effectuer une séparation ;

– *min_samples_leaf* : cet hyper-paramètre correspond au nombre d'instances minimum qu'une feuille doit contenir.

Afin de ne pas tester un trop grand nombre de combinaisons, une première modélisation sans appliquer de restrictions sur les hyper-paramètres est réalisée. Les valeurs proches des valeurs fournies par ce modèle sont alors testées pour déterminer les hyper-paramètres optimaux. Pour la modélisation par arbre de décision, les résultats peuvent être résumés de la manière suivante :

Hyper paramètre	**Domaine de définition**	**Valeur initiale**	**Intervalle**	**Valeur optimale**
max_depth	$\mathbb{N}^*$	23	$[\![5, 23]\!]$	11
min_samples_split	$\mathbb{N}^*$	2	$[\![1, 15]\!]$	6
min_samples_leaf	$\mathbb{N}^*$	1	$[\![1, 15]\!]$	1

TABLE 3.1 – Résumé de l'optimisation du modèle *CART*

3.2.2.3 Étape 3 : Indicateurs de performance

Dans la continuité de la méthode *Grid Search*, il est nécessaire de déterminer des métriques de validation qui permettront d'évaluer si un modèle est plus performant qu'un autre. Une métrique de validation est une mesure qui quantifie l'écart entre les données observées et les données prédites. Étant donné que cette étude concerne un problème de classification, la matrice de confusion sera à la base de différentes métriques.

Matrice de confusion : Dans un problème de classification binaire, les données prédites peuvent être de quatre types différents. Ceux-ci sont résumés dans la matrice de confusion suivante :

		Classe réelle observée	
		Non résiliation : 0	Résiliation : 1
Classe prédite	Non résiliation : 0	Vrais négatifs (VN)	Faux négatifs (FN)
	Résiliation : 1	Faux positifs (FP)	Vrais positifs (VP)

FIGURE 3.4 – Matrice de confusion

Cette matrice est obtenue sur la base de test. Dans le cadre de la modélisation par arbre de décision, la matrice de confusion obtenue est la suivante :

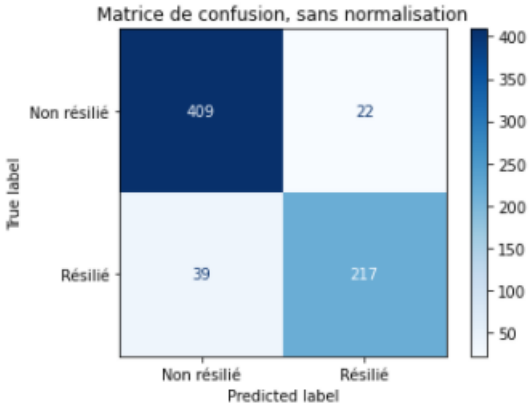


FIGURE 3.5 – Matrice de confusion pour le modèle par arbre de décision

De cette matrice de confusion, trois indicateurs peuvent être déduits.

Accuracy score : ce score détermine le pourcentage de correspondances exactes entre les données observées et les données prédites. Il s'obtient de la manière suivante :

$$Accuracy = \frac{VN + VP}{VN + VP + FN + FP}$$

Cette métrique donne une information globale sur la performance d'un modèle. Toutefois, elle ne permet pas de rentrer dans le détail des prédictions. En effet, dans le cadre d'une étude comme celle considérée, où il existe un fort déséquilibre entre les deux classes à prédire, un modèle qui affecte toutes les données comme appartenant à la classe majoritaire aura alors, au global, un bon *accuracy score*. S'agissant de la modélisation par arbre de décision, l'*accuracy score* obtenu est de 91,12 %. A première vue, ce score semble bon, il est toutefois nécessaire de l'affiner et notamment d'observer si les prêts qui devraient être étiquetés comme résiliés par le modèle le sont correctement. Pour ce faire, deux autres métriques permettent d'affiner ce résultat.

Precision score : ce score détermine la capacité du modèle à ne pas affecter une valeur positive à une donnée qui devrait être négative. Il s'obtient de la manière suivante :

$$Precision = \frac{VP}{VP + FP}$$

Cette métrique permet d'avoir une idée de la confiance qu'il est possible d'accorder à une donnée décrite comme résiliée. Plus le *precision score* est proche de 1, plus il est probable qu'un prêt décrit comme résilié le soit bien en réalité. Dans le cadre de la modélisation par arbre de décision, le *precision score* obtenu est de 90,79 %. Ce score révèle une réelle capacité du modèle à bien identifier les prêts qui vont être résiliés. Cependant, un *precision score* plus proche de 0 ne permettrait pas d'arriver aux mêmes conclusions. Certes, il est

possible que tous les prêts qui devraient être résiliés soient bien relevés par le modèle, mais il est également possible que celui-ci en relève aussi un grand nombre qui ne devraient pas l'être.

Recall score : ce score détermine la capacité d'un modèle à affecter une valeur positive à tous les éléments qui doivent l'être. Il s'obtient de la manière suivante :

$$Recall = \frac{VP}{VP + FN}$$

La résiliation ne représente qu'une petite partie des portefeuilles en assurance emprunteur. Il est donc important que les modèles qui sont entraînés soient en mesure de détecter correctement un prêt qui sera résilié quand ils en voient un. Pour la modélisation par arbre de décision, le *recall score* obtenu est de 84,77 %. Ce score montre la bonne capacité du modèle à identifier les prêts qui doivent être résiliés.

L'étude individuelle du *precision score* et du *recall score* n'est pas forcément des plus pertinente. Une étude conjointe de ces deux métriques permet d'arriver à des conclusions beaucoup plus convaincantes quant à la performance d'un modèle. Les scores de 90,79 % et 84,77 %, respectivement pour le *precision score* et le *recall score*, permettent d'arriver à la conclusion que, en plus de correctement identifier les prêts qui vont être résiliés, le modèle ne fait que peu d'erreur en n'identifiant que peu de prêts comme résiliés, alors qu'ils ne le sont pas. Dans l'hypothèse où seulement une seule des deux métriques serait bonne, un arbitrage devrait être effectué en fonction de l'objet de l'étude. D'un côté, le modèle identifiera un grand nombre d'instances comme positives, quitte à en identifier certaines qui ne devraient pas l'être. De l'autre, le modèle identifiera peu d'instances comme positives mais celles-ci seront très certainement correctement identifiées.

Des indicateurs complémentaires peuvent également apporter une information sur la performance globale du modèle.

Courbe ROC et AUC ROC : la courbe ROC (*Receiver Operating Characteristic*) trace les valeurs des taux de vrais positifs et taux de faux positifs pour différents seuils de classification.

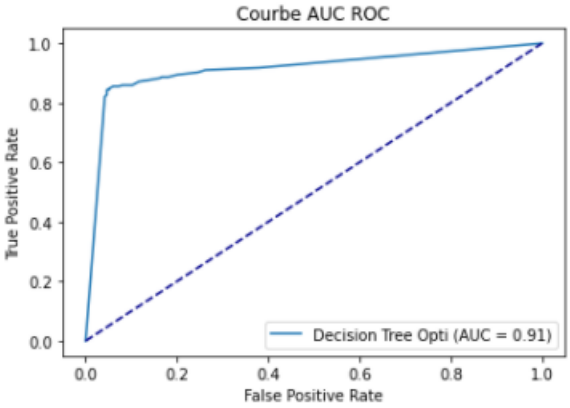


FIGURE 3.6 – Courbe ROC et AUC

L'AUC (*Area Under Curve*) correspond alors à l'aire sous la courbe ROC. Cette métrique donne une mesure agregée des performances pour tous les seuils de classification possibles. Comprise entre 0 et 1, elle est à mettre en comparaison avec le cas où les données seraient affectées aléatoirement aux deux classes. Dans ce cas, la courbe ROC correspond au segment allant du point (0,0) au point (1,1) et l'AUC vaut 50 %. Pour la modélisation par arbre de décision, l'AUC ROC a pour valeur 91%. En comparaison avec la *baseline* à 50%, cette valeur renforce l'intérêt pour cette méthode qui semble bien prédire les taux de résiliation.

Dans le cas de données déséquilibrées, ce qui est le cas dans cette étude, il peut être également utile d'utiliser la coubre P-R (*Precision - Recall*) associée à l'AUC qui en découle.

Courbe Precision-Recall et AUC P-R : le principe de fonctionnement de cette courbe est le même que la courbe ROC à l'exception près que ce sont le *precision score* et le *recall score* qui sont comparés pour différents seuils de classification. De la même manière, l'AUC correspond à l'aire sous la courbe P-R ainsi obtenue. Ici, l'AUC n'est pas à mettre en comparaison avec une *baseline* de 50% comme pour l'AUC ROC mais avec une *baseline* égale au taux de résiliation sur la base de données initiale (environ 20%).

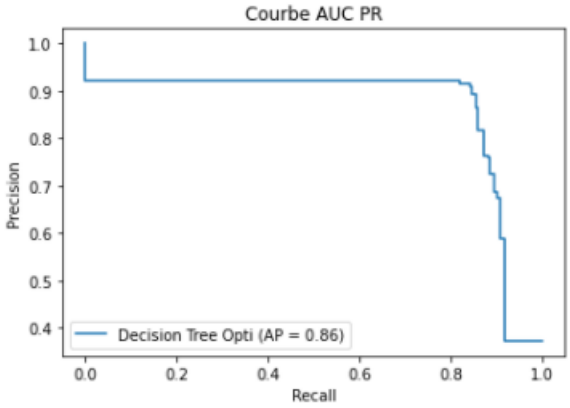


FIGURE 3.7 – Courbe PR et AUC

3.2.2.4 Etape 4 : Calcul et prolongement des taux de résiliation

Une fois la variable *Résiliation* prédite par le modèle, les taux de résiliation en fonction de l'ancienneté sont obtenus selon la formule suivante :

$$\text{Taux de résiliation}(i) = \frac{\text{nombre de prêts résiliés en } i}{\text{nombre de prêts dont Ancienneté an supérieur à } i}$$

Ensuite, la prolongation des taux jusqu'à 30 ans est obtenue à l'aide de la fonction de survie de Weibull, caractérisée par la formule suivante :

$$\widehat{q_n} = e^{-\lambda \cdot n^{\alpha}}$$

Avec :

- $\lambda = -\frac{ln(\widehat{q_n})}{n^{\alpha}}$ calculé au point de raccordement ;
- α paramètre permettant d'ajuster le degré de prudence avec lequel les taux sont prolongés.

Une présentation des taux de résiliation calculés sera réalisée dans le dernier chapitre de ce mémoire.

3.2.3 Limites de la méthode

Les arbres de décision présentent toutefois des limites non négligeables. En effet, le principe même de la méthode CART peut entraîner, dans certains cas, des arbres extrêmement complexes. Cela a pour conséquence de favoriser le sur-apprentissage : l'arbre obtenu risque de moins bien classer les données futures. Néanmoins, ce point peut être corrigé à l'aide des

hyper-paramètres au moment de la phase d'optimisation. Ce sujet sera traité dans la partie suivante de mise en pratique.

Les arbres de décision sont également sensibles aux variations dans la base de données. L'apprentissage sur des bases d'entraînement différentes peut produire des arbres aux prédictions différentes. Il est donc nécessaire de porter son attention sur la robustesse de ce modèle d'autant que nous sommes en présence d'une base de données où les deux classes sont fortement déséquilibrées : il y a une majorité de prêts non résiliés.

3.3 *Random Forest*

Dans cette partie, la modélisation des taux de résiliation va être abordée à l'aide de la méthode de *Random Forest*. Comme pour la partie précédente sur la modélisation par arbre de décision, les résultats obtenus seront explicités une fois les principes théoriques de la méthode présentés. Les limites de la méthode seront également abordées.

3.3.1 Principes théoriques

L'algorithme de forêts aléatoires, plus communément appelé *Random Forest* appartient à la famille des agrégations de modèles [Léo Breiman, 2001]. Il s'appuie sur les arbres de décisions étudiés dans la partie précédente en essayant de réduire au maximum la variance de ceux-ci. Le terme de *bagging*, contraction de *bootstrap aggregation*, est plus généralement utilisé. Dans le cadre du *Random Forest*, il s'agit même d'un double *bagging* à la fois sur les données et sur les variables explicatives.

L'algorithme de *Random Forest* est une technique qui consiste à construire N arbres à partir de N sous-ensembles du jeu de données initial. Ces sous-ensembles sont obtenus à l'aide d'un tirage aléatoire avec remise sur le jeu de données initial. Un tirage aléatoire est également réalisé sur les variables explicatives. En effet, si nous notons m le nombre de variables explicatives, chaque arbre créé dans le *Random Forest* sera entraîné sur $\sqrt{m}$ variables tirées aléatoirement. Une fois les arbres entraînés, la classe prédite sera celle présente en majorité dans les sorties des arbres.

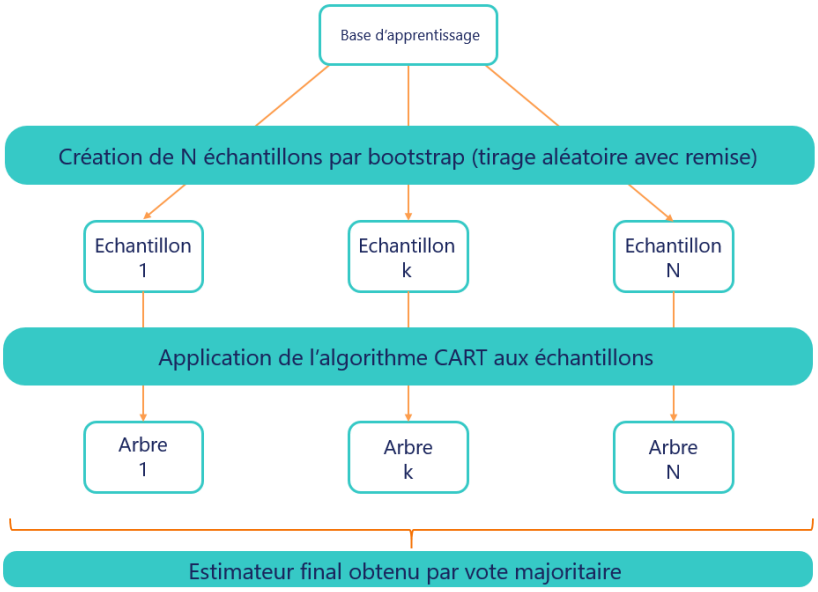


FIGURE 3.8 – Fonctionnement de l'algorithme de Random Forest

Pour optimiser ce modèle, il s'agit donc de travailler sur les hyper-paramètres du modèle CART que nous avons vus précédemment mais aussi sur les hyper-paramètres propres au *Random Forest* comme le nombre d'arbres créés.

3.3.2 Mise en pratique

L'algorithme de *Random Forest* se fondant sur la construction d'un grand nombre d'arbres, il est nécessaire d'accorder une attention encore plus importante à l'optimisation des hyper-paramètres. En effet, en plus de ceux du modèle CART, viennent s'ajouter les hyper-paramètres relatifs aux nombres d'arbres et de variables explicatives utilisées pour construire chaque arbre. Le temps de calcul étant un paramètre à prendre en compte lors de la construction de ce type de modèle, nous limiterons volontairement le nombre de combinaisons d'hyper paramètres que nous testerons. S'agissant de la démarche pour optimiser l'algorithme, nous adoptons un raisonnement similaire à celui des arbres de décision. Après avoir entraîné un modèle sans restriction particulière, nous récupérons les paramètres afin de concentrer nos recherches autour de ces valeurs.

Pour l'optimisation du *Random Forest*, les hyper-paramètres optimisés sont les suivants :

- *n_estimators* : cet hyper-paramètre correspond au nombre d'arbres créés ;
- *max_depth* : cet hyper-paramètre correspond à la profondeur maximale que peut avoir l'arbre ;
- *min_samples_split* : cet hyper-paramètre correspond au nombre d'instance minimum qu'un noeud doit avoir pour effectuer une séparation ;
- *min_samples_leaf* : cet hyper-paramètre correspond au nombre d'instances minimum qu'une feuille doit contenir ;
- *bootstrap* : cet hyper-paramètre indique si des échantillons bootstrappés sont utilisés pour entraîner les arbres. S'il prend la valeur *False*, la base de données d'entraînement complète est utilisée pour les entraîner.

Après optimisation par la méthode *Grid Search*, les résultats obtenus peuvent être résumés dans le tableau suivant :

Hyper paramètre	**Domaine de définition**	**Valeur initiale**	**Intervalle**	**Valeur optimale**
$n\_estimators$	$\mathbb{N}^*$	500	$[400, 600, 800, 1000, 1200, 1400]$	800
max_depth	$\mathbb{N}^*$	15	$[5, 10, 11, 12, 13, 14, 15]$	15
$min\_samples\_split$	$\mathbb{N}^*$	2	$[5, 10, 20]$	5
$min\_samples\_leaf$	$\mathbb{N}^*$	1	$[1, 2, 4, 6, 8, 10]$	1
$bootstrap$	$[True, False]$	$True$	$[True, False]$	$False$

TABLE 3.2 – Résumé de l'optimisation du modèle *Random Forest*

L'optimisation du modèle permet d'obtenir la matrice de confusion suivante :

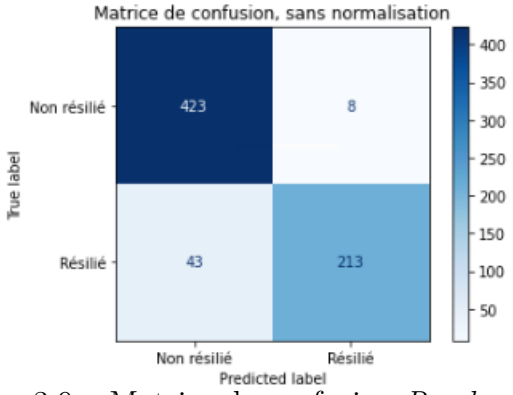


FIGURE 3.9 – Matrice de confusion *Random Forest*

De cette matrice, les métriques suivantes peuvent être déduites :

Accuracy score	*Precision score*	*Recall score*
92.58%	96.38%	83.20%

TABLE 3.3 – Métriques de performance après optimisation du *Random Forest*

Du point de vue des métriques d'*accuracy score* et de *precision score*, la modélisation par *Random Forest* est plus performante. Avec une *precision score* de 96.38%, ce modèle montre une réelle capacité à ne pas étiqueter comme résiliés les prêts qui ne le sont pas. Toutefois, cette performance se fait au dépend du *recall score* s'élevant à 83,20%. Ce score indique que le modèle identifie mal les prêts qui devraient être étiquetés comme résiliés. Ce phénomène aura pour conséquence de sous-estimer les taux de résiliation par rapport à ce qui se passe réellement.

3.3.3 Limites de la méthode

La structure même de l'algorithme de *Random Forest* permet de limiter le sur-apprentissage via la construction d'un grand nombre d'arbres. Toutefois, plus ces derniers sont nombreux, plus le temps de calcul nécessaire à l'entraînement de l'algorithme est long. De plus, cette multiplication du nombre d'arbres rend le fonctionnement de l'algorithme plus opaque pour l'utilisateur : nous parlons alors de boîte noire.

3.4 XGBoost

Dans cette partie, la modélisation des taux de résiliation va être abordée à l'aide de la méthode *XGBoost*. Le plan suivi sera le même que pour les parties précédentes à savoir les principes théoriques puis les résultats pratiques obtenus et enfin les limites liées à la méthode.

3.4.1 Principes théoriques

L'algorithme d'*eXtreme Gradient Boosting*, plus communément appelé *XGBoost*, appartient à la famille des méthodes ensemblistes [Chen and Guestrin, 2015]. Tout comme la méthode *Random Forest* étudiée précédemment, le modèle *XGBoost* va s'entraîner sur des sous parties diverses de la base d'entraînement puis faire voter ces sous-modèles pour prendre la décision finale. En plus de bénéficier d'un traitement parallélisé optimisant les temps de calcul, cette méthode fait appel au *boosting*. Le principe du *boosting* est de partir d'un modèle avec peu de performance prédictive et de l'améliorer au fur et à mesure en donnant plus de poids aux instances difficiles à prédire. Ainsi, le modèle améliore sa capacité prédictive en apprenant de ses erreurs.

En pratique, l'algorithme *XGBoost* fonctionne de la manière suivante :

1. Un premier arbre est entraîné sur une portion de la base d'entraînement ;
2. Les prédictions sur la base de test sont sauvegardées ;
3. Les prédictions sur la base d'entraînement permettent d'identifier les instances mal classées. L'algorithme affecte alors un poids plus important à ces instances ;
4. Le processus est répété jusqu'à convergence des résultats ;
5. L'estimateur final est obtenu par vote majoritaire à partir des prédictions issues des différents arbres créés.

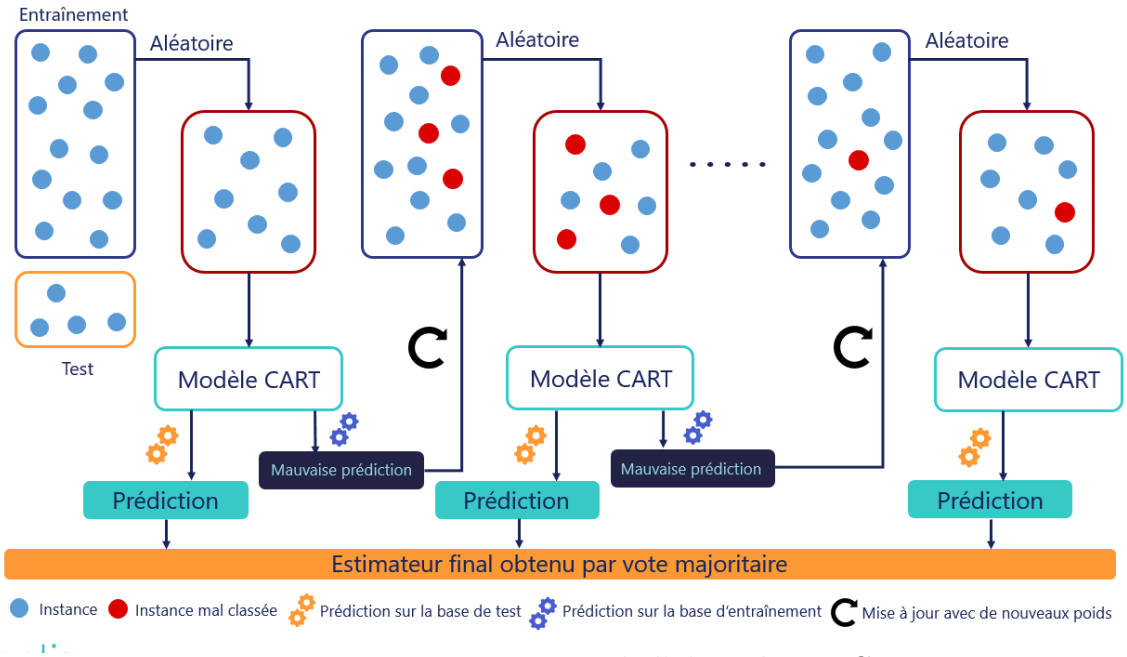


FIGURE 3.10 – Fonctionnement de l'algorithme XGBoost

3.4.2 Mise en pratique

L'algorithme *XGBoost* est réputé pour sa grande flexibilité fournissant de très bons résultats. Le revers de cette flexibilité est le nombre important d'hyper-paramètres supplémentaires à ajuster par rapport aux méthodes CART et *Random Forest*. De plus, même si cet algorithme est implanté dans les librairies de manière à optimiser le temps de calcul, il est nécessaire de ne pas tester trop de combinaisons d'hyper-paramètres lors de la phase d'optimisation. Par conséquent, de manière similaire aux deux méthodes prédécentes, la méthode d'optimisation par *Grid Search* sera utilisée sur un nombre restreint d'hyper-paramètres.

Pour l'optimisation du *XGBoost*, les hyper-paramètres optimisés sont les suivants :

- *learning_rate :* contrôle l'ajout de poids aux instances mal prédites d'un arbre à l'autre. Plus le *learning rate* est faible, plus le poids ajouté aux instances mal classées d'un arbre à l'autre est faible ;
- *max_depth :* correspond à la profondeur maximale que les arbres créés peuvent avoir ;
- *min_child_weight :* correspond à la valeur minimale de la somme des poids que doit contenir un nouvel arbre créé ;
- *gamma :* correspond à la valeur minimale de réduction de la fonction d'erreur requise pour réaliser une nouvelle partition d'un noeud dans les arbres ;

– *colsample_bytree :* correspond à la proportion de variables explicatives qui seront utilisées à chaque nouvel arbre créé.

Après optimisation par la méthode *Grid Search*, les résultats obtenus peuvent être résumés dans le tableau suivant :

Hyper paramètre	**Domaine de définition**	**Valeur initiale**	**Intervalle**	**Valeur optimale**
learning_rate	$[0, 1]$	0.3	$[0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.5, 1]$	0.25
max_depth	$\mathbb{N}^*$	6	$[6, 10, 12, 15]$	15
min_child_weight	$\mathbb{R}^*$	1	$[1, 3, 5, 7]$	1
gamma	$[0, 1]$	0	$[0, 0.1, 0.2, 0.3]$	0.1
colsample_bytree	$[0, 1]$	1	$[0.3, 0.4, 0.5, 0.7]$	0.7

TABLE 3.4 – Résumé de l'optimisation du modèle *XGBoost*

L'optimisation du modèle permet d'obtenir la matrice de confusion suivante :

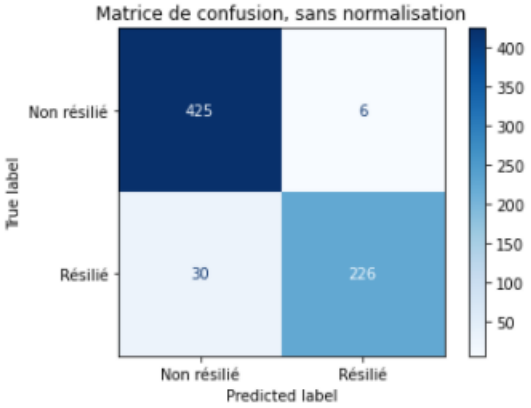


FIGURE 3.11 – Matrice de confusion *XGBoost*

De cette matrice, les métriques suivantes peuvent être déduites :

Accuracy score	*Precision score*	*Recall score*
94.47%	95.42%	89.45%

TABLE 3.5 – Métriques de performance après optimisation du *Random Forest*

La modélisation par *XGBoost* est également plus performante que la méthode CART (*Accuracy score* : 91,12%, *Precision score* : 90,79 %, *Recall score* : 84,77%). Concernant la comparaison avec le modèle par *Random Forest*, le modèle *XGBoost* semble plus performant du point de vue de l'*accuracy score* et du *recall score*. Des trois modèles testés, ce modèle

est donc celui qui identifie le mieux les prêts qui doivent être résiliés. C'est notamment l'augmentation du *recall score* qui permet d'expliquer l'augmentation de l'*accuracy score*. Toutefois, les résultats obtenus pour les modélisations par *Random Forest* et *XGBoost* sont sensiblement les mêmes. En l'état, il n'est pas possible de conclure à une supériorité d'un modèle ou de l'autre. L'objet de la partie suivante sera donc de tester la robustesse de ces modèles afin de définir lequel est le plus performant.

3.4.3 Limites de la méthode

Tout comme la méthode *Random Forest*, la structure de la méthode *XGBoost* permet de limiter le sur-apprentissage. Néanmoins, la grande flexibilité du modèle entraînant un temps de calcul important, il est nécessaire d'accorder une attention particulière aux nombres d'hyperparamètres optimisés. La contrainte temporelle devient alors un facteur limitant à toute étude impliquant un modèle de la sorte. Du point de vue interprétabilité de la méthode, le modèle *XGBoost* présente les mêmes défauts que la méthode *Random Forest* : la complexité de l'algorithme implique un phénomène de boîte noire rendant complexe l'interprétation des modèles étudiés.

3.5 Comparaison des modèles

3.5.1 Comparaison des indicateurs

L'entraînement d'un modèle sur une seule séparation de la base de données initiale ne permet pas de garantir un résultat optimal. En effet, il se peut que l'aléatoire fasse séparer la base de données en une situation particulièrement avantageuse pour les modèles. Ces derniers pourraient très bien être de mauvais prédicteurs pour tout autre séparation aléatoire. Pour pallier ce problème, il est nécessaire d'entraîner les modèles sur un grand nombre de séparations différentes de la base de données initiale. Il sera alors possible d'obtenir une densité des différentes métriques calculées.

En pratique, pour réaliser N entraînements d'un modèle, il faut créer N séparations de la base de données initiale à partir de graines aléatoires toutes différentes. Pour chaque séparation, le modèle est entraîné en suivant les trois premières étapes explicitées précédemment. A chaque itération de graine aléatoire, les métriques de performance sont stockées. Pour chaque métrique, il existe alors des vecteurs de taille N permettant d'obtenir des informations importantes sur la répartition de ces dernières.

Ces tests sur la robustesse des modèles sont nécessaires pour obtenir des résultats satisfaisants. Toutefois, cette démarche est fortement coûteuse en temps de calcul. Il faut donc

accorder une attention particulière au nombre de séparations réalisées. Ce point est particulièrement vrai sur des méthodes comme le *Random Forest* nécessitant structurellement un long temps de calcul. Pour cette étude, le nombre de séparations est fixé arbitrairement à 100 pour chaque modèle.

Les densités obtenues pour les divers modèles sont les suivantes :

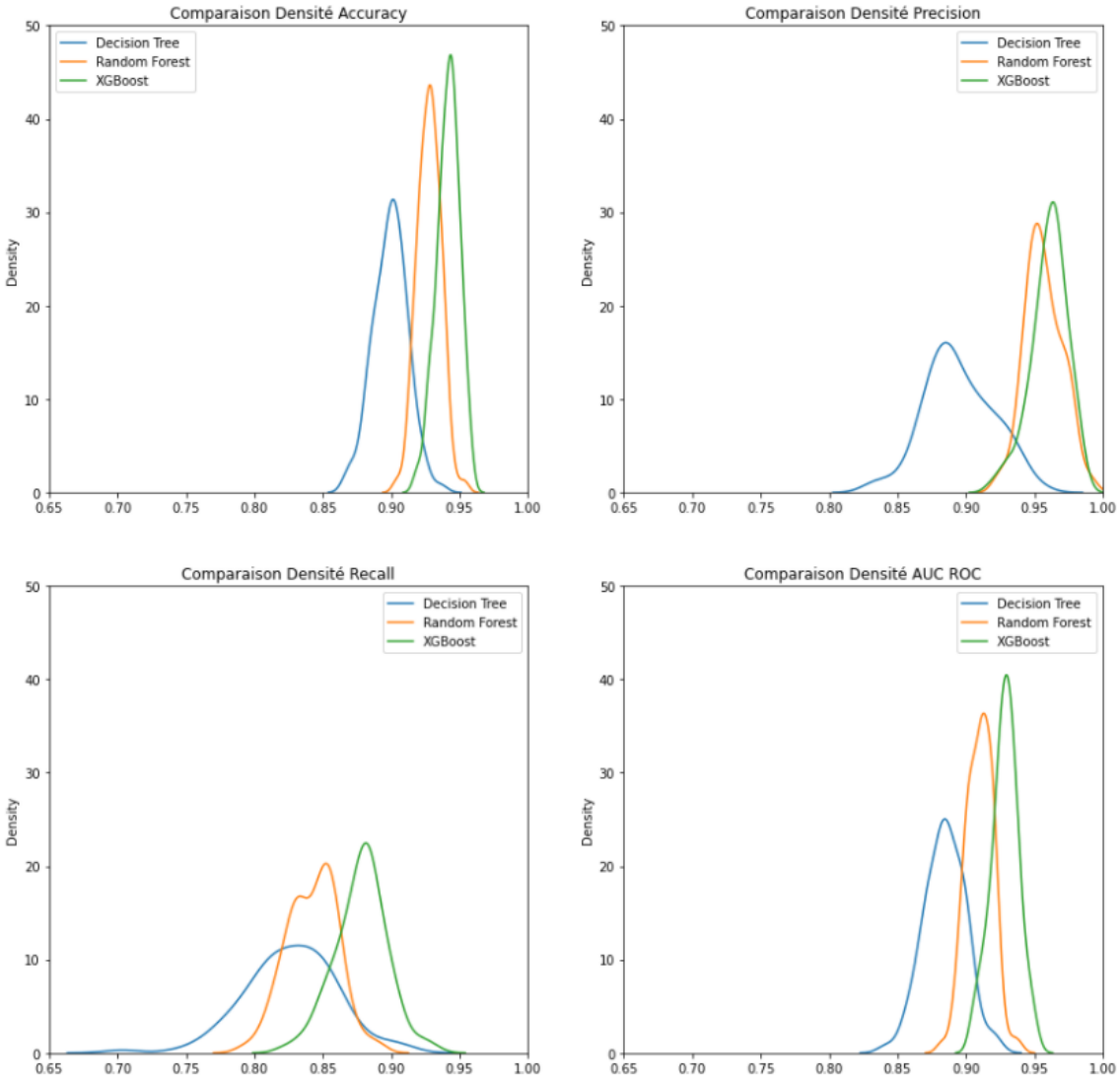


FIGURE 3.12 – Densité des différentes métriques de performance

Plusieurs conclusions peuvent être tirées de ces densités. D'une part, d'un point de vue performance, la méthode par *XGBoost* est celle qui convient le mieux pour prédire la résiliation des prêts. C'est également cette méthode qui présente les plus faibles taux de faux positifs et de faux négatifs (du fait des *precision score* et *recall score* élevés). Au contraire, la modélisation par arbre de décision est la méthode qui présente les moins bonnes performances

au global. Ces observations confirment les limites mentionnées précédemment, bien que les arbres de décision présentent une facilité de compréhension, ils n'ont pas une bonne capacité prédictive en comparaison des méthodes de *Random Forest* et de *XGBoost*.

D'autre part, ces densités fournissent des informations concernant la sensibilité des modèles à la séparation aléatoire de la base de données initiale en base d'entraînement et base de test. Qu'importe la métrique considérée, la méthode par arbre de décision est le modèle présentant le plus grand écart-type. Au contraire, les méthodes de *Random Forest* et de *XGBoost* présentent de faibles écarts-types pour toutes ces métriques. Un écart-type faible est représentatif d'un modèle peu soumis à l'aléatoire de la séparation de la base initiale en base d'entraînement et base de test.

En considérant les résultats observés, si un choix devait être fait parmi ces trois méthodes selon un critère de performance pure, la décision devrait s'orienter vers la méthode de *XGBoost*. Toutefois, si un arbitrage doit être réalisé entre performance et temps de calcul, la méthode par arbre de décision est la plus pertinente. Cette méthode présente des performances proches de celles des méthodes les plus performantes tout en offrant un temps de calcul rapide.

3.5.2 Comparaison des métriques agrégées

Les métriques agrégées que sont l'AUC ROC et l'AUC P-R permettent également de tirer des conclusions sur la performance des modèles étudiés. En considérant le modèle par arbre de décision comme le modèle de base pour les comparaisons, il est possible d'évaluer la relation entre gain de performance et gain en précision sur les prédictions.

Les densités pour les deux métriques évoquées sont les suivantes :

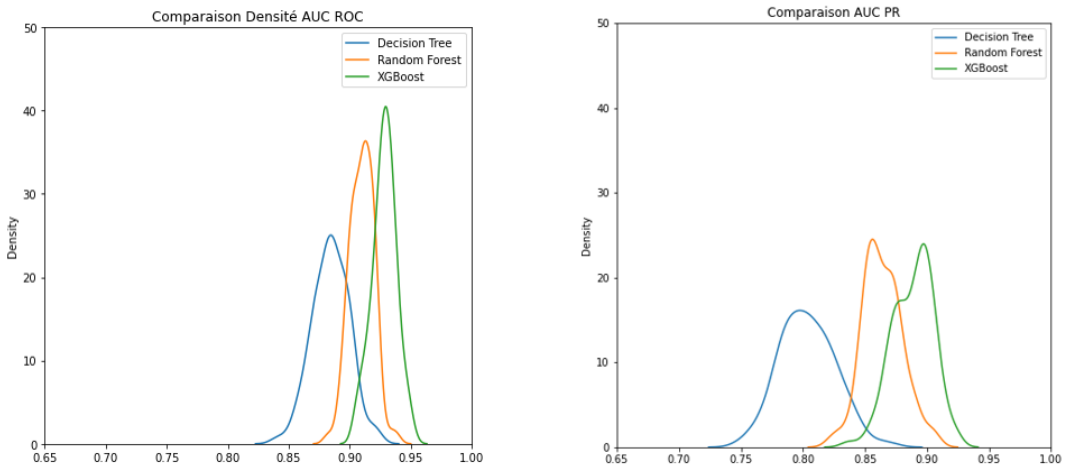


FIGURE 3.13 – Densités obtenues pour l'AUC ROC et l'AUC P-R

Ces métriques permettent de souligner un point important. Bien que le gain en performance globale, représenté par l'AUC ROC, soit minime, il est possible que le gain en précision soit important. Les valeurs obtenues pour chaque modèle optimisé viennent confirmer ces observations :

Modèle	AUC ROC	Gain de performance	AUC P-R	Gain de précision
Decision Tree	89,83 %		82,64 %	
Random Forest	90,67 %	+ 0,94 %	86,45 %	+ 4,61 %
XGBoost	93,45 %	+ 4,03 %	89,28 %	+ 8,03 %

TABLE 3.6 – Comparaison des métriques agrégées

En comparant l'AUC ROC obtenue pour le modèle par arbre de décision avec sa densité associée, nous constatons que celle-ci fait partie des valeurs les plus élevées que le modèle peut fournir. Même dans ce cas de figure idéal, les modèles *Random Forest* et *XGBoost* présentent des performances bien supérieures. Toutefois, ici aussi, une distinction peut être faite entre ces deux méthodes. Bien que le gain en précision soit non négligeable pour le *Random Forest* par rapport à l'arbre de décision, le gain en performance globale est moindre. Ce n'est pas le cas pour le modèle *XGBoost* qui gagne à la fois en performance globale et en précision. A temps de calcul égal, il est donc plus intéressant, d'un point de vue performance, de se tourner vers le modèle *XGBoost*.

Chapitre 4

Application des méthodes

L'objectif de ce dernier chapitre est de comparer les lois de résiliation obtenues avec les différentes méthodes. A ce stade de l'étude, les taux de résiliation à notre disposition sont les suivants :

- taux de résiliation obtenus à l'aide de l'estimateur de Kaplan-Meier et le lissage de Whittaker-Henderson ;
- taux de résiliation obtenus à l'aide d'un arbre de décision ;
- taux de résiliation obtenus à l'aide d'un modèle *Random Forest* ;
- taux de résiliation obtenus à l'aide d'un modèle XGBoost.

Ces différentes lois permettront de mesurer l'impact du choix de la méthode sur la rentabilité d'un contrat d'assurance emprunteur, via le calcul d'un indicateur de rentabilité. De plus, la notion de censure est à prendre en compte pour les méthodes de *Machine Learning* et sera étudiée au travers du modèle de *Random Survival Forest*. Les méthodes de *Machine Learning* apportent également des fonctionnalités supplémentaires, telles que l'influence des variables, permettant une approche différente dans la construction des taux de résiliation. Enfin, ce chapitre sera l'occasion de présenter les limites propres à cette étude et d'évoquer les travaux complémentaires qui pourraient être menés afin d'approfondir le sujet.

4.1 Comparaison des taux de résiliation

4.1.1 Mise en commun des taux de résiliation obtenus

Après avoir déterminé les taux de résiliation selon la méthode actuarielle et selon les méthodes de *Machine Learning*, il est possible de comparer les taux de résiliation entre eux. Graphiquement, les taux de résiliation obtenus sont les suivants :

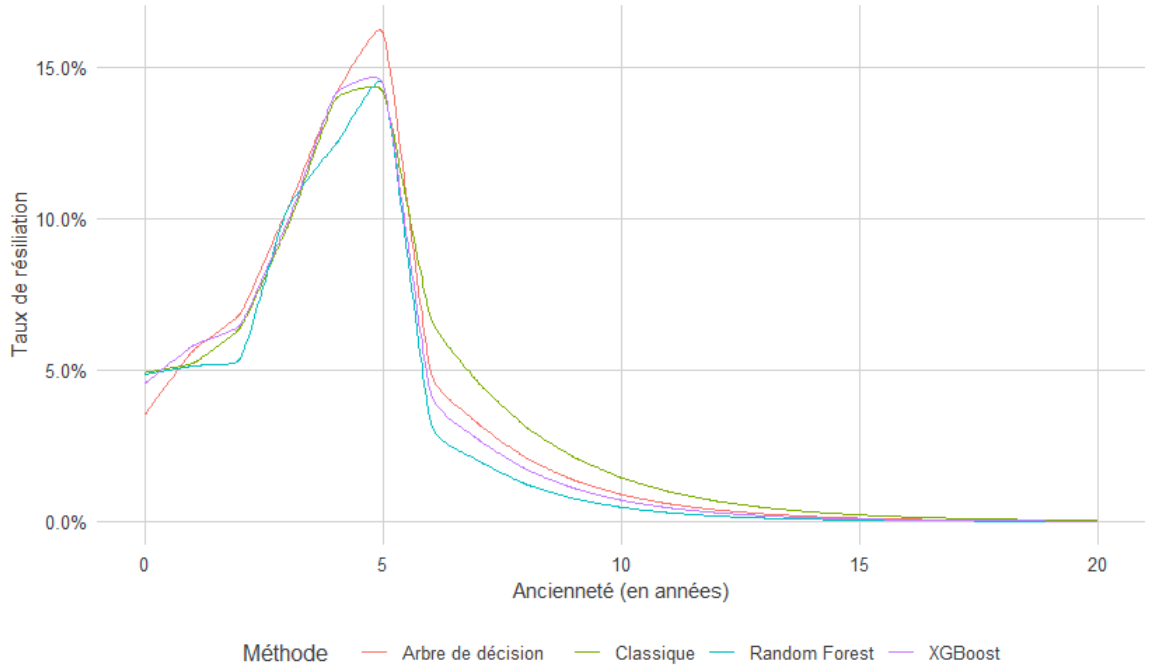


FIGURE 4.1 – Taux de résiliation pour les différentes méthodes

Bien que les taux présentent une allure similaire, il n'est pas possible de déterminer graphiquement si l'une des méthodes se détache des autres d'un point de vue économique. C'est pour cette raison que la suite de cette étude consistera à déterminer un indicateur de performance sur un portefeuille de contrats d'assurance emprunteur. Cet indicateur sera calculé sur un portefeuille en situation de *run-off*, c'est-à-dire un portefeuille sans nouvel entrant et qui est projeté jusqu'à son extinction.

4.1.2 Indicateur retenu

L'indicateur de rentabilité qui permettra d'évaluer l'impact du choix de la méthode retenue pour la construction des taux de résiliation est la **Valeur Actuelle Probable** (VAP) des résultats financiers nets sur toute la vie du produit. La VAP se calcule de la manière suivante :

$$VAP = \sum_{k=1}^{n} \frac{Resultats\ nets_k}{(1+i)^k}$$

Où i est le taux d'actualisation retenu.

Selon la garantie étudiée, le taux d'actualisation utilisé sera différent. Les taux employés seront ceux préconisés par l'article A132-1 du Code des Assurances à savoir :

– le taux technique vie sera appliqué pour l'actualisation des flux propres au décès ;

– le taux technique non-vie sera appliqué pour l'actualisation des flux propres à l'arrêt de travail.

La méthode employant l'estimateur de Kaplan-Meier et le lissage de Whittaker-Henderson étant la plus communément utilisée, c'est celle qui servira de base de comparaison avec les autres méthodes. Plus précisément, c'est le ratio I des VAP du modèle de base et du modèle à comparer qui permettra de mesurer l'évolution entre les deux méthodes. I s'écrit alors de la manière suivante :

$$I = \frac{VAP_{méthode\ comparée}}{VAP_{méthode\ actuarielle}} - 1$$

Toutes les étapes nécessaires au paramétrage et à l'obtention de ces VAP sont détaillées dans le mémoire *Etude d'une opportunité de diversification par l'ORSA : application à l'assurance emprunteur* [PEREZ, 2017].

4.1.3 Tarification

Avant de pouvoir procéder au calcul des VAP, il est nécessaire de passer par la tarification d'un produit emprunteur. Ainsi, les tarifs utilisés sont ceux obtenus dans le cadre du mémoire évoqué ci-dessus. Ces tarifs seront intégrés au *business plan* permettant de projeter les résultats nets et de calculer les VAP.

De plus, en cohérence avec le portefeuille étudié, les tarifs utilisés ont été calculés selon une tarification au capital initial. Ce choix est également justifié par le fait que le marché est dominé par les contrats collectifs qui se basent sur ce principe [FFA, 2020]. Les formules utilisées pour la tarification sont celles détaillées la partie 1.2.

4.1.4 Résultats obtenus

Pour évaluer l'impact des méthodes de *Machine Learning* au travers de la VAP, le choix a été fait d'utiliser les informations de la base de données utilisées pour l'étude. Pour ce faire, deux scénarios ont été envisagés avec les caractéristiques suivantes :

Scénario 1 : Sans segmentation du portefeuille

Pour ce premier scénario, il a été décidé de conserver les statistiques moyennes de la base de données sans segmentation utilisée pour l'étude. Les hypothèses faites sont donc les suivantes :

- **Âge à la souscription :** 35 ans ;
- **Durée d'emprunt :** 16 ans ;
- **Montant emprunté :** 150 000 € ;
- **Taux d'emprunt :** 1,10 % ;
- **Nombre de prêts en $t = 0$:** 3400 prêts.

Scénario 2 : Avec segmentation du portefeuille en classes d'âge

Pour ce second scénario, il a été décidé d'affiner les paramètres en prenant en compte la segmentation mentionnée dans la section 2.2.3, à savoir une segmentation en trois classes d'âges. Les hypothèses faites sur ces classes sont les suivantes :

Classe d'âge	**[20 ; 30 [**	**[30 ; 40 [**	**[40 ; 75]**
Âge à la souscription	27 ans	34 ans	45 ans
Durée d'emprunt	18 ans	16 ans	14 ans
Montant emprunté	110 000 €	170 000 €	155 000 €
Taux d'emprunt	1,10 %	1,10 %	1,10 %
Nombre de prêts en $t = 0$	800 prêts	2000 prêts	600 prêts

TABLE 4.1 – Hypothèses du scénario 2

Pour les deux scénarios, les taux techniques utilisés pour l'actualisation des flux sont ceux du 31 juillet 2021 à savoir 0,00 % pour le taux technique vie et -0,04 % pour le taux technique non-vie.

Une première comparaison, du point de vue du ratio I, peut être réalisée en prenant comme méthode de comparaison les taux de résiliation calculés à partir des données observées. Toutefois, compte tenu de la fenêtre d'observation restreinte, cette comparaison ne peut se faire que sur les premières années d'ancienneté. Les résultats obtenus sont les suivants :

Méthode considérée	Scénario 1	Scénario 2
Observée		
Actuarielle	+ 1,41 %	+ 0,95 %
Arbre de décision	+ 1,56 %	+ 1,02 %
Random Forest	+ 2,34 %	+ 1,83 %
XGBoost	+ 1,29 %	+ 0,76 %

TABLE 4.2 – Comparaison des méthodes avec les données observées pour les scénarios 1 et 2

Bien que les taux de résiliation obtenus sur les premières années d'anciennetés avec les données observés soient plus prudents que toutes les autres méthodes, ils ne permettent pas d'aboutir à une loi de résiliation applicable dans le cadre d'une projection de portefeuille. C'est pourquoi, notre choix s'est porté sur la comparaison des méthodes de *Machine Learning* avec la méthode actuarielle classique. Les résultats obtenus pour les deux scénarios envisagés peuvent se résumer dans le tableau suivant :

Méthode considérée	Scénario 1	Scénario 2
Actuarielle		
Arbre de décision	+ 0,46 %	+ 0,49 %
Random Forest	+ 2,37 %	+ 1,87 %
XGBoost	+ 0,70 %	+ 0,49 %

TABLE 4.3 – Comparaison des méthodes pour les scénarios 1 et 2

Quel que soit le scénario, l'approche selon la méthode actuarielle classique est la plus prudente. Toutefois, la méthode d'estimation des taux de résiliation par *Random Forest* présente l'augmentation de la VAP la plus importante. La notion de censure, très présente dans la base de données utilisée pour cette étude, n'étant pas prise en compte dans les méthodes de *Machine Learning*, nous nous proposons d'essayer d'expliquer cette augmentation en s'appuyant sur une extension du modèle de *Random Forest* prenant en compte ce principe : le modèle de *Random Survival Forest*.

4.2 *Random Survival Forest*

4.2.1 Présentation générale du modèle

Le modèle de forêt de survie aléatoire, plus communément appelé *Random Survival Forest*, permet d'introduire la notion de censure dans les modèles de *Machine Learning*. En effet, pour les méthodes détaillées au chapitre 3, cette dernière n'est pas prise en compte.

La méthode *Random Survival Forest* est une extension de la méthode *Random Forest* de régression appliquée à l'analyse de survie [Hemant Ishwaran and Lauer, 2008]. Elle est fondée sur le même principe de double *bagging* à la fois sur les données et les variables explicatives. Mais, à la différence du modèle de régression classique, le critère de séparation pour la division de chaque noeud est un critère incluant la durée de survie et la censure. De plus, au lieu de retourner une classe ou un score comme pour les méthodes de classification et de régression, la méthode de *Random Survival Forest* retour une fonction de survie estimée à l'aide de l'estimateur de Nelson-Aalen pour chaque feuille terminale.

4.2.2 Optimisation du modèle

Pour optimiser le modèle lors de la phase d'entraînement, la méthode *GridSearchCV* a été utilisée. Dans le cadre du modèle de *Random Survival Forest*, les hyper-paramètres à optimiser sont les mêmes que ceux mentionnés à la section 3.3.2 et les valeurs optimales obtenues sont résumées dans le tableau suivant :

Hyper paramètre	**Domaine de définition**	**Intervalle**	**Valeur optimale**
$n\_estimators$	$\mathbb{N}^*$	$[800, 1000, 1200, 1400, 1600]$	1400
max_depth	$\mathbb{N}^*$	$[6, 8, 12, 14, 16, 18, 20]$	18
$min\_samples\_split$	$\mathbb{N}^*$	$[3, 4, 5, 10]$	3
$min\_samples\_leaf$	$\mathbb{N}^*$	$[1, 2, 4, 5]$	1

TABLE 4.4 – Résumé de l'optimisation du modèle *Random Survival Forest*

Concernant la métrique de performance utilisée pour optimiser le modèle, la méthode de *Random Survival Forest* emploie l'indice de concordance. Cet indice estime la probabilité, dans une sélection aléatoire de données, que la donnée qui a la plus courte durée de survie est celle qui la moins bonne estimation. Plus concrètement, cet indice est souvent interprété comme une probabilité de mauvaise classification. Dans le cadre de ce modèle, les indices de concordance obtenus sur la base de *train* et la base de *test* sont les suivants :

Métrique	Base de *train*	Base de *test*
Indice de concordance	98,87 %	74,75 %

TABLE 4.5 – Indices de concordance obtenus après optimisation du modèle de *Random Survival Forest*

L'indice de concordance est à mettre en comparaison avec une *baseline* à 50%. En effet, comme pour l'interprétation de la courbe ROC, un score de 50 % correspond aux résultats d'un modèle où toutes les prédictions sont déterminées de manière aléatoire. Les scores obtenus permettent donc de montrer que le modèle optimisé est performant en comparaison de cette *baseline*.

4.2.3 Résultats obtenus

Une fois le modèle optimisé, chaque ligne de prêt dans la base de test permet d'obtenir une fonction de survie. Ainsi, il est alors possible de déterminer une fonction de survie globale en calculant la moyenne de toutes les fonctions de survie obtenues. De la même manière, il est possible d'obtenir les intervalles de confiance à 95 % pour cette même fonction de survie. Ainsi, cette dernière prend la forme suivante :

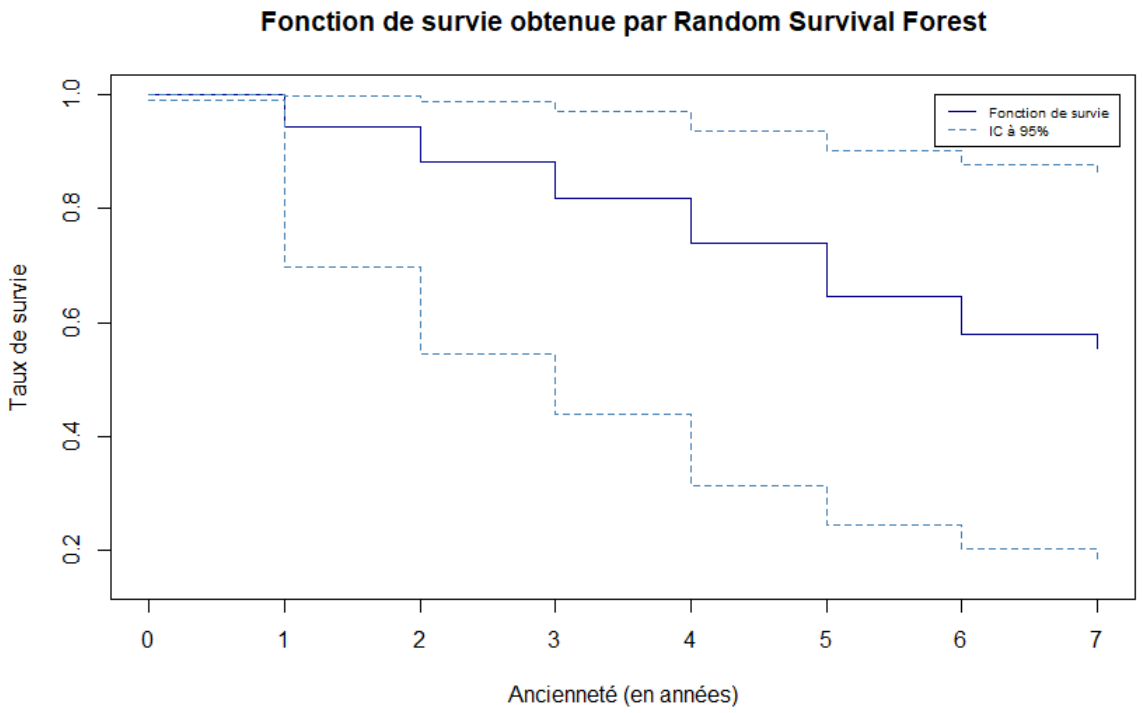


FIGURE 4.2 – Fonction de survie et intervalles de confiance à 95 % pour le modèle de *Random Survival Forest*

Une fois la fonction de survie obtenue, il est possible de passer aux taux de résiliation selon la même formule que pour la modélisation actuarielle classique. Ensuite, la loi est prolongée à l'aide la fonction de survie de Weibull. Une fois la loi de résiliation complète obtenue, les ratios I sont calculés selon les deux scénarios évoqués afin de mesurer l'impact de l'ajout de la censure dans les méthodes de *Machine Learning*. Les résultats obtenus sont les suivants :

Ratio I	Scénario 1	Scénario 2
Random Forest	+ 2,37 %	+ 1,87 %
Random Survival Forest	+ 0,50 %	+ 0,13 %

TABLE 4.6 – Comparaison du modèle de *Random Survival Forest* pour les scénarios 1 et 2

Ces résultats montrent, dans le cadre de notre étude, que le passage à des modèles de *Machine Learning* intégrant la censure permet d'obtenir une loi de résiliation plus prudente. Toutefois, cette approche additionnelle ne permet pas d'atteindre le degré de prudence obtenu avec la méthode actuarielle classique.

4.3 Apport du *Machine Learning*

4.3.1 Influence des variables

Les méthodes de *Machine Learning* permettent, en plus de modéliser les taux de résiliation des contrats d'assurance emprunteur, de déterminer l'importance des différentes variables dans l'explication du phénomène de résiliation et donc dans la modélisation de ces taux.

Pour rappel, les méthodes de classification utilisées dans cette étude utilisent le critère de Gini comme critère d'optimisation. Ainsi, pour les méthodes par arbre de décision et par *Random Forest*, l'importance d'une variable correspond à son influence dans la diminution du critère de Gini lors de la phase d'entraînement du modèle. Pour ce qui est du F-score, celui-ci correspond au nombre de fois qu'une variable est séparée en 2 lors de la phase d'entraînement du modèle XGBoost.

Les quatre variables explicatives qui ressortent des différents modèles peuvent être résumées dans le tableau suivant :

Modèle	Arbre de décision		*Random Forest*		XGBoost	
Variable	Rang	Importance	Rang	Importance	Rang	F-score
Ancienneté	1ère	24,21 %	1ère	27,24 %	3ème	426
Année de souscription	2ème	17,28 %	2ème	19,21 %	4ème	255
Montant emprunté	3ème	8,74 %	3ème	7,79 %	1ère	853
Année de naissance	4ème	6,97 %	4ème	6,14 %	2ème	500

TABLE 4.7 – Tableau récapitulatif de l'influence des variables sur les taux de résiliation

A partir de ces informations, il est possible d'envisager l'expression des taux de résiliation des contrats d'assurance emprunteur selon d'autres variables. Une étude des corrélations entre ces quatre variables permet alors d'apporter des éléments complémentaires pour le choix de la variable qui servira de segmentation supplémentaire.

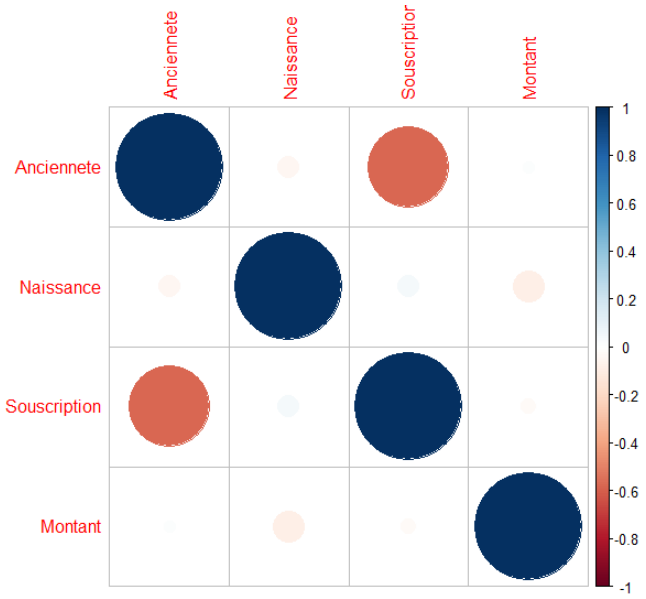


FIGURE 4.3 – Matrice de corrélation des quatre variables ayant le plus d'influence sur les taux de résiliation

Bien que la variable représentant l'année de souscription soit la deuxième variable la plus importante pour calculer les taux de résiliation, la corrélation importante qui existe entre cette variable et l'ancienneté du prêt ne la rend pas pertinente pour une nouvelle segmentation. De plus, même si l'année de naissance ne présente pas de corrélation importante avec l'ancienneté, une telle segmentation a déjà été réalisée dans le chapitre 2. Par conséquent, notre choix s'est porté sur une segmentation selon le montant emprunté.

4.3.1.1 Taux de résiliation en fonction du montant emprunté

Tout comme la segmentation en classe d'âge du chapitre 2, le choix d'une segmentation en trois parties a été retenu. Ce choix provient du raisonnement selon laquelle les prêts appartiennent à l'une des catégories suivantes :

- **catégorie 1 :** le prêt réalisé est d'un faible montant et vient en complément d'un apport ou d'un autre prêt ;
- **catégorie 2 :** le prêt réalisé est d'un montant intermédiaire et correspond à l'achat d'un bien de valeur moyenne ;
- **catégorie 3 :** le prêt réalisé est d'un montant important est correspond au prêt principal lors de l'achat d'un bien de grande valeur.

Afin de respecter la cohérence des hypothèses mentionnées ci-dessus il a été décidé de fixer les seuils entre les catégories 1 et 2 et les catégories 2 et 3 à respectivement 100 000 € et 200 000 €.

La formule utilisée pour déterminer les taux de résiliation selon le montant est la suivante :

$$\text{Taux\_de\_résiliation}(i,j) = \frac{\text{nombre de prêts } j \text{ résiliés en } i}{\text{nombre de prêts } j \text{ dont Ancienneté an supérieur } i}$$

Où :

- i correspond à l'ancienneté du prêt ;
- j correspond à la catégorie de prêt à laquelle appartient le prêt.

Les taux de résiliation obtenus selon le montant emprunté peuvent être résumés dans le graphique suivant :

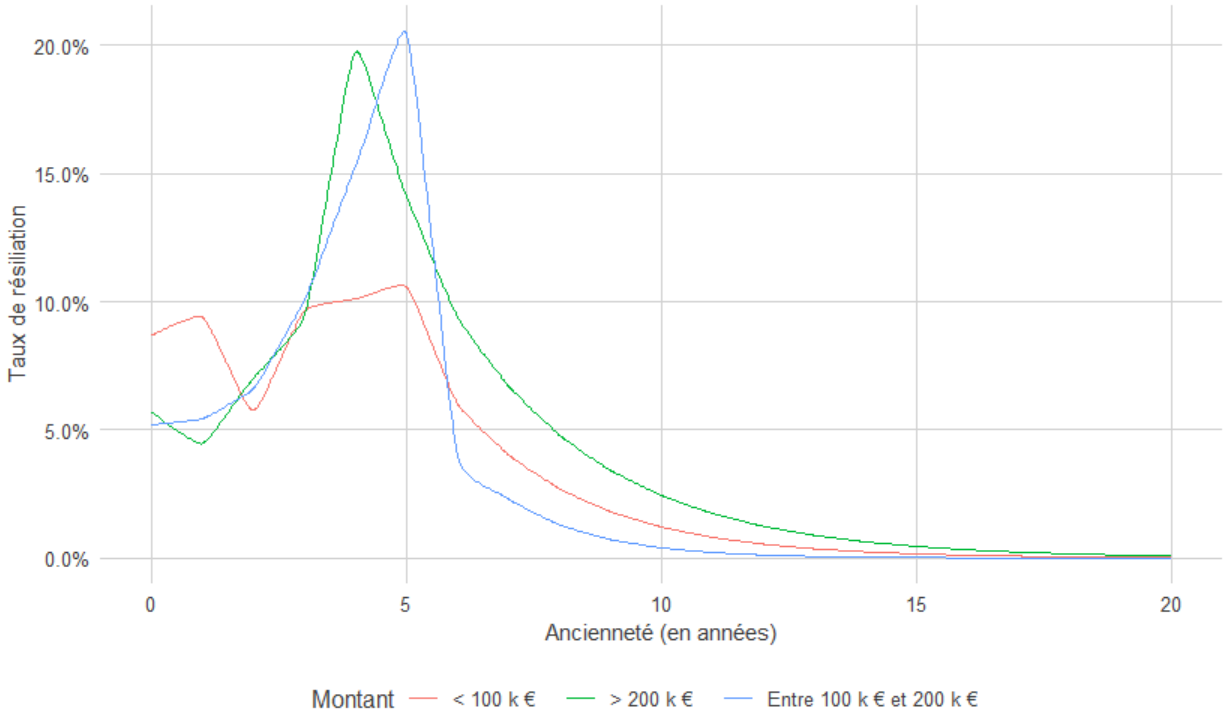


FIGURE 4.4 – Taux de résiliation segmentés par montant emprunté

De ces taux de résiliation se dégagent deux tendances. D'un côté, les prêts dont le montant emprunté est inférieur à 100 000 €. De l'autre, les prêts dont le montant emprunté est supérieur à 100 000 €. Au-delà des irrégularités observées sur le graphique et qui sont probablement dues à une faible exposition, il est possible de faire des conjectures sur le comportements des assurés. Pour les prêts dont le montant emprunté est inférieur à 100 000 €, les taux de résiliation ne dépassent pas les 10 %. Cela traduit le faible intérêt que peuvent avoir les assurés à résilier leurs contrats pour ces prêts. En effet, la prime d'assurance étant

proportionnelle au montant emprunté, l'intérêt financier est moindre lorsque le montant emprunté est faible. Au contraire, pour les prêts aux montant plus importants, la résiliation peut se justifier par une plus grande possibilité d'économie sur le montant de son assurance emprunteur.

4.4 Limites et ouvertures

Dans cette section, les difficultés rencontrées durant cette étude seront abordées. Ce sera également l'occasion d'évoquer les opportunités d'approches complémentaires offertes par cette étude.

4.4.1 Limites rencontrées lors de l'étude

La qualité des données occupe une place centrale dans les limites à mentionner. Pour mener à bien cette étude, de nombreux retraitements ont été nécessaires pour rendre la base de données de départ exploitable. Ces modifications entraînent automatiquement la mise en place d'hypothèses qui influencent les résultats obtenus. Outre les informations renseignées pour chaque variable, le nombre de variables est également un critère à prendre en compte. La présence de variables supplémentaires, comme le taux d'emprunt par exemple, aurait permis d'intégrer une composante supplémentaire. Au niveau des méthodes de *Machine Learning*, la présence de variables supplémentaires aurait également permis d'élargir le périmètre de l'étude voire d'améliorer l'optimisation des modèles.

La taille du jeu de données est également un sujet important dans les limites à aborder. Il est communément admis que plus une base de données est grande, plus les modèles ajustés sur celle-ci seront optimaux. Dans le cadre de cette étude, une taille plus importante du jeu de données aurait permis d'éviter les problèmes d'exposition rencontrés sur les anciennetés les plus grandes. De plus, une modélisation mensuelle ou bien une modélisation couvrant un plus grand nombre d'anciennetés aurait pu être rendue possible. Cette taille de la base de données est également un argument qui est entré en ligne de compte lors du choix des méthodes de modélisation des taux de résiliation. En effet, certains modèles de *Data Science*, notamment les modèles de *Deep Learning* avec les réseaux de neurones artificiels, sont réputés pour leur grande efficacité mais ceux-ci nécessitent une quantité de données en amont que cette étude ne permettait pas.

4.4.2 Ouvertures possibles

En supposant que les limites liées à la qualité et la quantité des données soient corrigées, ce mémoire pourrait être à la base d'études complémentaires. En disposant d'un portefeuille emprunteur conséquent, il serait possible de construire un outil permettant de suivre son évolution d'une année sur l'autre en utilisant les méthodes de *Machine Learning* et de *Deep Learning*. En prédisant les résiliations tout au long de la vie du portefeuille, cet outil pourrait potentiellement permettre d'optimiser la rentabilité du portefeuille.

De plus, toujours en supposant que les limites liées à la qualité et la quantité des données soient résolues, il serait possible de construire un outil de *scoring* à partir de méthodes de *Machine Learning* et de *Deep Learning*. En affectant un score à chaque prêt en fonction de ses caractéristiques (montant emprunté, durée du prêt, ...), cet outil permettrait de piloter le portefeuille en fonction de la politique de souscription de l'entreprise.

Conclusion

Les évolutions législatives engagées par le gouvernement ces dix dernières années au niveau de l'assurance emprunteur ont pour objectif d'ouvrir ce marché à la concurrence. Désormais, il est possible de résilier son contrat emprunteur sans aucune modification sur le prêt sous-jacent grâce à l'amendement Bourquin.

Ce mémoire nous a permis d'étudier ce phénomène de résiliation à l'aide de méthodes de *Machine Learning*. Ces méthodes, qui ont fait leurs preuves en assurance non-vie, sont une nouvelle manière d'aborder le risque de résiliation en assurance emprunteur.

Le processus d'optimisation des méthodes de *Machine Learning* apporte des premiers éléments quant à la pertinence des lois obtenues pour cette étude. Le modèle XGBoost semble être celui qui représente le plus fidèlement le risque de résiliation du portefeuille étudié avec des AUC ROC et AUC P-R respectivement égaux à 93,45 % et 89,28 %. La comparaison de la méthode actuarielle classique avec les méthodes de *Machine Learning* révèle que ces dernières sont toutes moins prudentes que la méthode actuarielle classique. Cependant, la prise en compte de la censure dans le cadre du modèle de *Random Survival Forest* permet de limiter cette augmentation de la VAP.

Toutefois, les méthodes de *Machine Learning* s'accompagnent de fonctionnalités permettant d'évaluer l'influence des variables dans la modélisation du risque de résiliation. Outre le fait que pour que cette étude le montant emprunté soit l'une des variables les plus influente dans la détermination des taux de résiliation, ces indicateurs permettent d'envisager de nouvelles manières de segmenter les taux de résiliation.

Concernant les limites, la qualité et la quantité des données sont des facteurs importants à prendre en considération. Un nombre d'informations plus important, notamment sur les caractéristiques des emprunteurs ou encore les taux d'emprunt, ainsi qu'un échantillon de prêts plus conséquent auraient permis d'envisager d'autres méthodes de modélisation ou bien d'affiner les résultats obtenus. Même si cette étude ne permet pas d'aboutir à une préférence pour les méthodes de *Machine Learning*, elle peut toutefois être à la base d'études

complémentaires. Celles-ci pourraient approfondir les méthodes déjà étudiées ou bien se concentrer sur la construction d'un outil prédictif du risque de résiliation.

Bibliographie

[Chen and Guestrin, 2015] Chen, T. and Guestrin, C. (2015). Xgboost : A scalable tree boosting system. Technical report, LearningSys.

[FFA, 2020] FFA (2020). L'assurance française - chiffres clés 2019. .

[France Info, 2019] France Info (2019). Crédit immobilier : quel est le profil type de l'emprunteur ? Repéré sur : *https://www.francetvinfo.fr/economie/immobilier/credit-immobilier-quel-est-le-profil-type-de-l-emprunteur_3166589.html*.

[Hemant Ishwaran and Lauer, 2008] Hemant Ishwaran, Udaya B. Kogalur, E. H. B. and Lauer, M. S. (2008). Random survival forest. Technical report, Institute of Mathematical Statistics.

[IGF, 2013] IGF (2013). Assurance emprunteur. Repéré sur : *http://www.igf.finances.gouv.fr/files/live/sites/igf/files/contributed/IGF%20internet/2.RapportsPublics/2013/2013-M-086.pdf*.

[Kaplan et Meier, 1958] Kaplan et Meier (1958). Nonparametric estimation from incomplete observations.

[Lee Giesecke, 1981] Lee Giesecke (1981). Use of the chi-square statistic to set whittaker-henderson smoothing coefficients.

[Léo Breiman, 1996] Léo Breiman (1996). Bagging predictors.

[Léo Breiman, 2001] Léo Breiman (2001). Machine learning.

[Meilleurtaux.com, 2019] Meilleurtaux.com (2019). Baisse du revenu moyen et de l'âge des emprunteurs immobiliers en 2018. Repéré sur : *https://www.meilleurtaux.com/credit-immobilier/actualites/2019-fevrier/baisse-du-revenu-moyen-et-de-l-age-des-emprunteurs-immobiliers-en-2018.html*.

[Ministère de l'Economie et des Finances, 2018] Ministère de l'Economie et des Finances (2018). AERAS - les points-clés. Repéré sur : *http://www.aeras-infos.fr/cms/sites/aeras/accueil/aeras-en-pratique/les-points-cles.html* [Consulté le 9 août 2020].

[PEREZ, 2017] PEREZ, L. (2017). Etude d'une opportunité de diversification par l'orsa : application à l'assurance emprunteur. Mémoire de master, DUAS.

Annexes

Annexe 1 : Extraits du Code des Assurances

Article L141-1

"Est un contrat d'assurance de groupe le contrat souscrit par une personne morale ou un chef d'entreprise en vue de l'adhésion d'un ensemble de personnes répondant à des conditions définies au contrat, pour la couverture des risques dépendant de la durée de la vie humaine, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque de chômage. Les adhérents doivent avoir un lien de même nature avec le souscripteur."

Article R321-1

"L'agrément administratif prévu par l'article L. 321-1 est accordé par l'Autorité de contrôle prudentiel et de résolution. Pour l'octroi de cet agrément, les opérations d'assurance sont classées en branches et sous-branches de la manière suivante :

1. Accidents (y compris les accidents de travail et les maladies professionnelles)

[...]

20. Vie-Décès : Toute opération comportant des engagements dont l'exécution dépend de la durée de la vie humaine autre que les activités visées aux branches 22,23 et 26."

Article R343-3

"1° Provision mathématique : différence entre les valeurs actuelles des engagements respectivement pris par l'assureur et par les assurés. Pour des contrats faisant intervenir une table de survie ou de mortalité, les montants des provisions mathématiques doivent inclure une estimation des frais futurs de gestion qui seront supportés par l'assureur pendant la période de couverture au-delà de la durée de paiement des primes ou de la date du prélèvement du capital constitutif ; l'estimation de ces frais est égale au montant des chargements de gestion prévus dans les conditions tarifaires de la prime ou du capital constitutif et destinés à couvrir les frais de gestion ;"

Article R331-6

"5° Provision pour risques croissants : provision pouvant être exigée, dans les conditions fixées par le décret prévu à l'avant-dernier alinéa de l'article R. 331-1, pour les opérations d'assurance contre les risques de maladie et d'invalidité et égale à la différence des valeurs actuelles des engagements respectivement pris par l'assureur et par les assurés ;"

Annexe 2 : Extrait du Code de la Consommation

Article L313-30

"Jusqu'à la signature par l'emprunteur de l'offre mentionnée à l'article L313-24, le prêteur ne peut pas refuser en garantie un autre contrat d'assurance dès lors que ce contrat présente un niveau de garantie équivalent au contrat d'assurance de groupe qu'il propose. Il en est de même lorsque l'emprunteur fait usage du droit de résiliation mentionné au premier alinéa de l'article L113-12-2 du code des assurance ou au deuxième alinéa de l'article L221-10 du code de la mutualité dans un délai de douze mois à compter de la signature de l'offre de prêt mentionnée à l'article L313-24 ou qu'il fait usage du droit de résiliation annuel mentionné au deuxième alinéa de l'article L113-12 du code des assurances ou premier alinéa de l'article L221-10 du code de la mutualité. Toute décision de refus doit être motivée."

Table des figures

Liste des tableaux