IRM | institut du risk management BY Institut A

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Fabrice TAMBOURAN

Titre : Apport des méthodes de machine learning sur la prédiction du risque de résiliation d'un portefeuille d'assurance vie épargne dans un contexte transactionnel

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

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise : Crédit Mutuel Arkéa

Nom :

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : Yann Choquet

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

Résumé

L’objectif de l’étude est de fournir une démarche alternative pour calibrer les lois de résiliation dans les portefeuilles d'assurance vie en Unité de Compte (UC) dans un contexte transactionnel, où les enjeux de valorisation sont importants bien que les informations soient souvent limitées.

L'étude exploite des méthodes de machine learning pour analyser les résiliations historiques et développer un modèle prédictif spécifique au portefeuille étudié, tenant compte des particularités comportementales de la clientèle. L’étude cherche à prédire les résiliations futures pour gagner en précision et ainsi valoriser le portefeuille de l’entité faisant l’objet de la transaction de manière plus fine et adaptée aux spécificités du portefeuille.

Dans un contexte transactionnel, cela évite l’application des lois de résiliation du repreneur, par manque d’information, pouvant introduire des biais car souvent peu représentatives de la clientèle du portefeuille cible et ne considérant pas la composante conjoncturelle pour les contrats UC car non imposée par le régulateur et pénalisante.

Malgré ses avantages, cette approche fait face à des défis, notamment la complexité d'application du machine learning dans un contexte transactionnel exigeant fiabilité et efficacité opérationnelle. Ces limitations soulignent l'importance de l'interprétabilité des modèles et leur validation par la fonction actuarielle.

Dans ce contexte, prédire spécifiquement le risque de résiliation du portefeuille cible représente un avantage stratégique crucial pour les acquéreurs potentiels, pouvant conduire à un avantage compétitif substantiel dans le processus transactionnel.

Summary

The objective of the study is to provide an alternative approach to calibrate surrender laws in unit-linked life insurance portfolios in a transactional context, where valuation stakes are high, despite often limited information.

The study leverages machine learning methods to analyse historical surrenders and develop a predictive model specific to the portfolio under study, considering the behavioural characteristics of the clientele. The study aims to obtain a more accurate information and more visibility to predict future surrenders to value the portfolio of the entity subject to the transaction more closely given the specificities of the portefolio.

In a transactional context, this avoids applying the acquirer's surrender laws, which, due to a lack of information, can introduce biases as they often do not represent the clientele of the target portfolio and do not consider the conjunctural component for unit-linked contracts, as it is not imposed by the regulator and is penalizing.

Despite its advantages, this approach faces challenges, particularly the complexity of applying machine learning in a transactional context that demands reliability and operational efficiency. These limitations highlight the importance of model interpretability and validation by the actuarial function.

In this context, accurately predicting the surrender risk of the target portfolio represents a crucial strategic advantage for potential acquirers, potentially leading to a substantial competitive edge in the transaction process.

Note de synthèse

En France, les assurances vie en Unités de Compte (UC) attirent de nombreux investisseurs grâce à leur potentiel de gains élevés et leur flexibilité, malgré l'absence de garantie sur le capital investi. Cette flexibilité permet aux souscripteurs de réagir rapidement aux variations du marché, en rachetant ou résiliant leurs contrats à tout moment.

Cependant, cela pose un défi pour les assureurs, nécessitant une gestion prudente et dynamique des fonds en UC pour maintenir la stabilité financière et la rentabilité. Les décisions de rachat ou de résiliation par les assurés peuvent impacter significativement les actifs sous gestion et les revenus des chargements de gestion. Les assureurs doivent aussi parfois liquider des actifs à des prix inférieurs pour satisfaire ces demandes.

Pour mitiger ces risques, ils réalisent des analyses comportementales détaillées et calibrent actuariellement les lois de résiliation des contrats pour anticiper les résiliations. Cela permet d'adapter leur gestion des actifs, maintenir une rentabilité stable et répondre efficacement aux exigences de liquidité des souscripteurs.

Dans les transactions de compagnies d'assurance, la valeur intrinsèque du portefeuille joue un rôle capital dans la détermination de la valeur de la cible, influençant fortement le prix que le repreneur serait disposé à payer. Cette valeur est directement affectée par la probabilité de résiliation des contrats. Un portefeuille relativement stable avec un faible taux de résiliation est davantage valorisé car il reflète des revenus futurs plus durables, ce qui en fait une acquisition plus sécurisée. Inversement, un taux élevé de résiliation signale une plus grande incertitude sur la durabilité des flux de revenus futurs et rend ainsi le portefeuille moins attrayant.

Lors d'un processus transactionnel, l'information mise à disposition des potentiels repreneurs est souvent limitée et les lois de résiliations ne sont en général pas communiquées durant la phase d'étude en amont de la transaction. Dans les faits, il est usuel que le potentiel repreneur utilise ses propres lois de résiliation sur portefeuille de la cible, en l'adaptant dans la mesure du possible, pour son exercice de valorisation. Cela génère un biais car la loi de résiliation du repreneur est souvent peu représentative de la clientèle du portefeuille cible, et peut ne pas comporter de composante conjoncturelle car non imposée par le régulateur et pénalisante pour les calculs d'engagements.

Dans ce contexte, l'objectif principal du mémoire est de fournir une démarche alternative pour un meilleur calibrage du risque de résiliation d'un portefeuille de contrats UC dans un contexte transactionnel. L'étude exploite des méthodes de machine learning pour analyser les données de résiliation et développer un modèle prédictif spécifique au portefeuille étudié, en tenant compte des particularités comportementales de la clientèle. Elle vise ainsi à proposer une méthode de calibrage des

lois de résiliation plus représentative que les lois internes de l'acquéreur potentiel, souvent appliquées par défaut en raison des restrictions d'information pendant la phase de due diligence et de valorisation avant la remise de l'offre ferme auprès des vendeurs.

Ce calibrage sur mesure se révèle être stratégique car il permettrait une évaluation plus fine du portefeuille cible par une meilleure estimation du risque de résiliation et il pourrait offrir un avantage concurrentiel significatif dans le processus de transaction.

Du fait de l'importance stratégique de ces prédictions et des enjeux financiers que cela représente pour les acquéreurs potentiels, le mémoire reconnaît également les limites que comporte cette approche dans un contexte transactionnel. Les limitations de l'approche proviennent de la qualité des informations disponibles et des difficultés à appliquer le machine learning dans un contexte transactionnel concret. En effet, dans ces situations, l'efficacité opérationnelle et la rapidité d'exécution sont cruciales en raison des délais resserrés pour l'étude. De plus, l'interprétabilité des modèles et leur approbation par la fonction actuarielle sont essentielles au regard des enjeux financiers.

L'étude se concentre sur l'analyse des données de résiliation et le développement d'un modèle prédictif qui vise non seulement à donner du confort sur un exercice de valorisation dans un contexte transactionnel, mais aussi à proposer une méthode alternative aux méthodes traditionnelles de calibrage des lois de résiliation afin d'améliorer la perception du risque de résiliation des portefeuilles d'assurance vie en UC prenant en compte les spécificités comportementales des clients face aux fluctuations économiques et financières.

Cette double perspective ouvre la voie à de nouvelles pratiques par l'intégration grandissante du machine learning dans le secteur des assurances, où la maîtrise de la donnée est au cœur de l'activité.

L'analyse des données

L'étude s'appuie sur le dataset 'eusavingULnoPSperYr', partie du package 'CASdatasets' élaboré par Arthur Charpentier et Christophe Dutang, qui compile des données sur des produits d'épargne en UC, spécifiquement des contrats Unit-Linked sans participation aux bénéfices, émanant d'assureurs anonymes de l'Union Européenne. Ce dataset est considéré comme représentatif d'un portefeuille standard d'une entreprise d'assurance vie en UC. Le dataset comprend des données relatives à la période de 1999 à 2007, qui englobe plusieurs phases importantes du cycle économique européen, marquées par des périodes de croissance et de récession, offrant un aperçu des tendances et comportements des souscripteurs sur près d'une décennie. Il est important de mentionner que le dataset fait état uniquement des entrées en relation et des résiliations entre 1999 et 2007.

L'idée est d'analyser ce dataset pour déceler les tendances comportementales des souscripteurs en réponse aux variations des indicateurs économiques et financiers, notamment les variations relatives de

l’Indice de Prix à la Consommation (CPI – Consumer Price Index) ou de l’indice boursier européen sur différentes périodes.

L'analyse des données constitue une étape importante de la démarche, avec pour but une compréhension visuelle des éléments influençant les décisions de résiliation et l’identification des anomalies pouvant biaiser les modèles.

La phase initiale de l’Analyse Exploratoire des Données (EDA) commence par une analyse visuelle via des graphiques et des visualisations diverses, essentielle pour repérer les tendances et anomalies, et pour identifier les attributs clés pour le futur modèle prédictif. Cette démarche vise à saisir les facteurs critiques dans le processus de résiliation.

Ensuite, nous réalisons une étude technique pour approfondir l'examen des caractéristiques du dataset, en analysant les distributions, corrélations et autres statistiques descriptives pour comprendre les interactions entre variables. L'EDA, incluant le traitement des valeurs manquantes et l'identification de valeurs aberrantes, préparera les données pour la modélisation, en veillant à la fiabilité et à l'efficacité du modèle.

La finalité est de structurer le dataset de manière à optimiser les performances du modèle prédictif, en choisissant judicieusement les variables et en peaufinant la préparation des données pour améliorer la précision des prédictions.

La construction du modèle prédictif

Le développement d'un modèle prédictif implique plusieurs phases clés du prétraitement des données à l'évaluation des modèles.

La première étape, le data preprocessing, transforme les données brutes en un format adapté à l'analyse, impliquant nettoyage, normalisation et encodage pour optimiser l'entrée dans les modèles de machine learning.

Ensuite, le choix stratégique des algorithmes permet d'évaluer divers modèles pour trouver l'outil le plus efficace pour notre prédiction. Parmi eux, la Régression Logistique pour la classification binaire, les Forêts Aléatoires pour leur capacité à traiter des relations complexes entre un nombre importants de variables explicatives, XGBoost (eXtreme Gradient Boosting) pour sa performance et sa capacité à apprendre de ses erreurs, et enfin, les Réseaux de Neurones pour leur aptitude à capturer des relations non linéaires.

Chaque modèle est évalué selon plusieurs critères de performance (précision, rappel, courbe ROC, matrice de confusion, interprétabilité).

Suite à cette évaluation, une phase d'optimisation est entreprise pour préciser les paramètres et améliorer les performances, choisissant finalement le modèle XGBoost pour sa capacité à fournir des prédictions précises et fiables, avec un meilleur équilibre entre les différents indicateurs de performance :

Modèle	Accuracy	Precision	Recall	F1-score	AUC
Régression Logistique	0.6244	0.25	0.002	0.0039	0.6787
Arbres de Décision	0.7706	0.3545	**0.3797**	0.3667	0.6166
Forêts Aléatoires	0.8281	0.5136	0.3209	**0.395**	0.7643
XGBoost	**0.8426**	**0.6073**	0.283	**0.3861**	**0.7873**
Réseaux de Neurones	**0.8414**	**0.6076**	0.2621	0.3662	**0.7772**

Voici les raisons principales de préconisation :

- **F1 Score :** Le modèle XGBoost a montré un F1 score élevé, indiquant un excellent équilibre entre la précision et le rappel. Cela signifie que le modèle est capable de prédire correctement un grand nombre de cas positifs tout en limitant le nombre de faux positifs.
- **Précision et Rappel :** Le modèle XGBoost a également démontré une précision et un rappel élevés, ce qui est crucial pour les applications où il est important de minimiser les erreurs de classification, comme dans la prévision du risque de résiliation.
- **AUC Score :** Avec un AUC (Area Under the ROC Curve) score très élevé, le modèle XGBoost a prouvé sa capacité à distinguer efficacement entre les classes positives et négatives. Un AUC élevé est un indicateur de la robustesse du modèle face à différents seuils de classification.
- **Matrices de Confusion :** Les matrices de confusion ont révélé que le modèle XGBoost a le meilleur équilibre entre les vrais positifs (TP) et les vrais négatifs (TN), avec un nombre réduit de faux positifs (FP) et de faux négatifs (FN). Cela indique une excellente capacité de classification correcte des instances.
- **Courbe ROC :** La courbe ROC du modèle XGBoost, se rapprochant plus de l'angle supérieur gauche, illustre sa supériorité en termes de taux de vrais positifs par rapport au taux de faux positifs. Cela témoigne de sa capacité à maximiser les vrais positifs tout en minimisant les faux positifs.

Le modèle XGBoost se distingue donc par sa performance globale supérieure sur l'ensemble des critères évalués. Au-delà des variables explicatives structurelles (âge de l'assuré et ancienneté des contrats) l'analyse des variables explicatives les plus influentes du modèle XGBoost fait ressortir deux variables clés :

- **CPI.relvar2yr :** indique l'évolution de l'inflation sur deux ans à travers le CPI. Les fluctuations du CPI affectent le pouvoir d'achat, influençant la capacité ou la disposition des consommateurs à maintenir le paiement de leurs primes, ce qui a un impact direct sur le risque de résiliation.

- **EUidx.relvar2yr** : représente l'évolution à long terme de l'indice boursier européen sur deux ans, signalant l'impact des tendances économiques sur les décisions de résiliation. Une variation notable peut révéler des ajustements dans l'environnement économique influençant la perception du risque et la stabilité financière des assurés.

La projection des comportements de résiliation et perspective de valorisation du portefeuille

Notre approche vise à estimer l'évolution future des résiliations au sein du portefeuille d'assurance vie en UC, pour évaluer la dégradation progressive du portefeuille, c'est-à-dire la réduction du nombre de contrats actifs sans prendre en compte de nouveaux contrats. Nous présumons l'absence de versements additionnels ou de nouveaux contrats dans nos projections pour mener notre analyse. Le dataset présentant des données historiques de 1999 à 2007, nous nous focalisons sur les contrats actifs à la fin de cette période comme base pour notre valorisation.

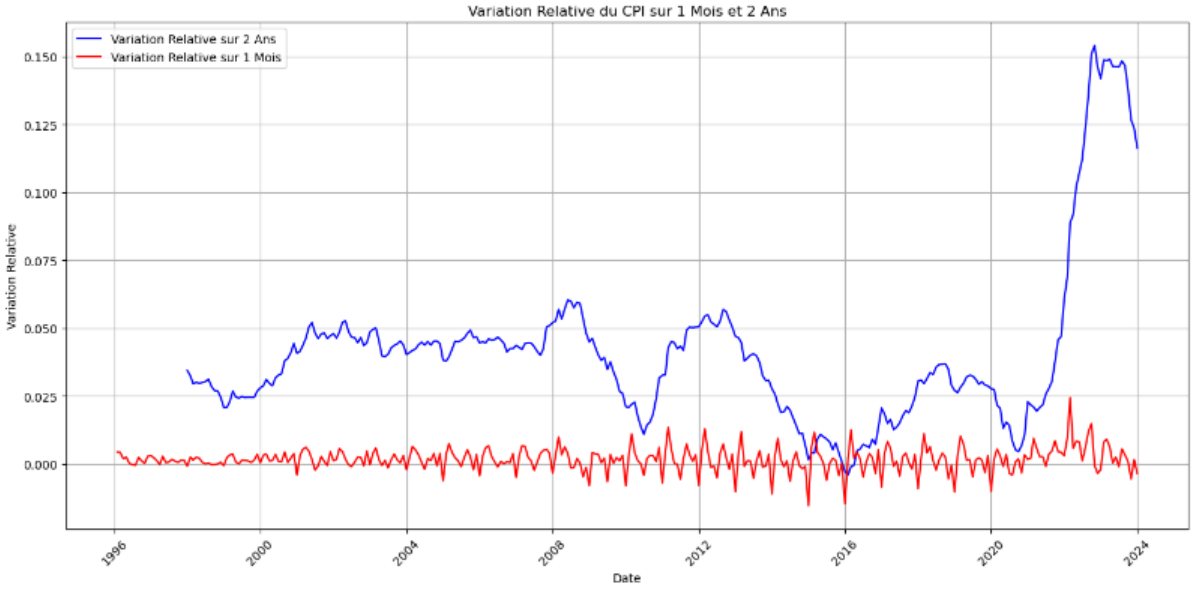
La suite de la démarche consiste à alimenter le modèle initialement entrainé sur les données historiques (ici XGBoost) d'un ou de plusieurs scénarios de l'environnement économique sur un horizon de projection pertinent pour établir l'écoulement du portefeuille. Nous procédons à la projection des variables économiques et financières les plus influentes de notre modèle, à savoir le CPI et l'indice boursier européen (EUidx), ainsi que des caractéristiques propres aux assurés, pour prévoir leur impact sur les décisions de résiliation.

Nous proposons plusieurs approches pour la projection de ces variables économiques et financières, selon un scénario déterministe, selon des séries temporelles qui permettent de capter les cycles économiques, ou selon des méthodes stochastiques. Les assureurs disposant en général de Générateurs de Scénarios Economiques, celui-ci est préconisé pour créer ces projections stochastiques des variables explicatives qui serviront comme données d'entrée dans le modèle de prédiction.

Les données dont nous disposons nous contraignent à réaliser la valorisation de ce portefeuille à fin 2007. Au lieu de générer des projections pour les variables explicatives, nous avons opté pour une approche simplifiée que nous pensons plus efficiente en utilisant les données économiques et financières réelles entre 2008 et 2023. A noter que cette démarche peut être suivie à la date de valorisation souhaitée, sous réserve de disposer des données jusqu'à cette date de valorisation. Il convient ensuite d'utiliser le Générateur de Scénarios Economiques pour projeter les variables explicatives comme données d'entrée du modèle. Dans nos cas, nous nous appuyons sur les évolutions historiques de l'indice Euro Stoxx 50, représentant les performances des grandes entreprises en zone euro, et du CPI, qui mesure l'inflation.

La valorisation d'un portefeuille d'assurance vie épargne nécessite en général un horizon de projection d'au moins 40 ans et en général entre 50 ans et 70 ans. Pour des raisons de simplification et d'efficience de notre étude, nous nous limitons à une période de 15 ans, utilisant exclusivement des données réelles

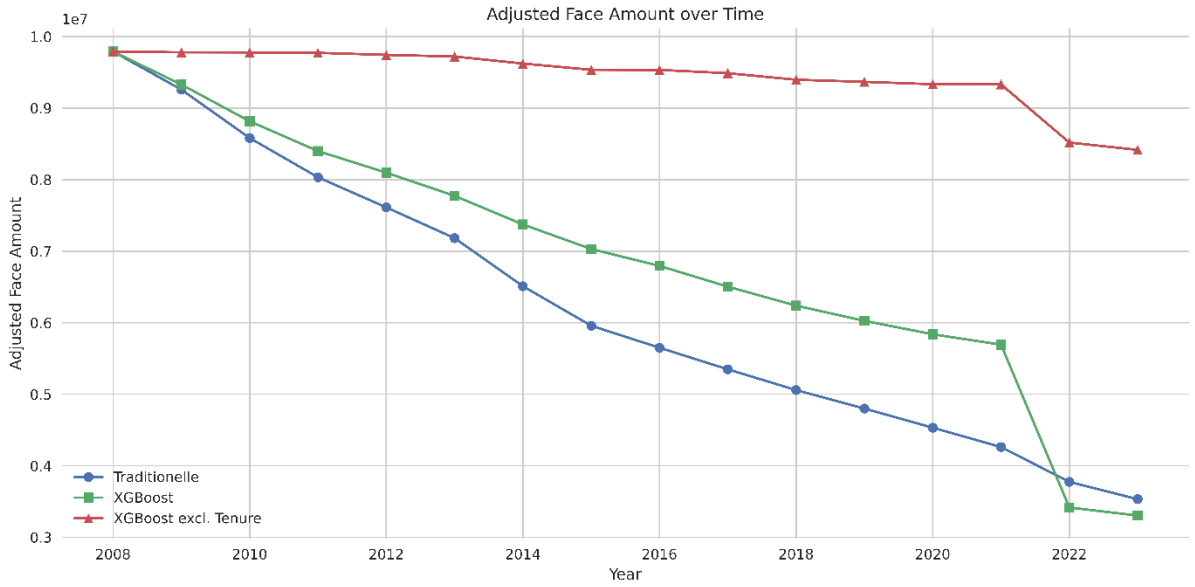
pour nos variables, ce qui couvre au moins un cycle économique et fait apparaître les impacts du Covid. Cette durée est suffisamment étendue pour permettre une analyse intéressante des tendances, tout en restant assez concise pour éviter les incertitudes trop lointaines et les complexités analytiques excessives dans le cadre de l'étude.



Variation relative sur 2 ans (bleu) et 1 mois (rouge) de l'indice CPI (Consumer Price Index)

Le modèle XGBoost entraîné est ensuite utilisé pour estimer les résiliations futures à partir d'un scénario déterministe de variables explicatives (CPI et Euro Stoxx 50) sur l'horizon 2008-2023.

A des fins de comparaison, nous avons simulé l'écoulement du portefeuille de contrats selon la méthode traditionnelle consistant à reprendre simplement les taux de résiliations historiques du portefeuille.



Ecoulement du portefeuille - Méthode Traditionnelle vs. Méthode de Machine Learning XGBoost (et à titre illustratif XGBoost excl. ancienneté)

La comparaison des écoulements montre que la méthode traditionnelle, basée sur l'évolution intrinsèque des caractéristiques du portefeuille (ancienneté et âge des assurés), présente un écoulement stable et quasi linéaire. En revanche, la méthode XGBoost révèle une chute significative de contrats en réponse au pic d'inflation, un facteur ignoré par la méthode traditionnelle. Cette dernière ne capture pas l'effet du pic inflationniste sur le comportement de résiliation des assurés.

En définitive, les résultats sont similaires : la méthode traditionnelle prévoit 64% de résiliations sur 15 ans, tandis que la méthode XGBoost en estime 66%, dont 14% attribués au pic inflationniste pendant le Covid. Sans cet effet, la méthode XGBoost prévoirait environ 50-55% de résiliations. L'explication de ce décalage est complexe en raison de l'aspect « boîte noire » des modèles de machine learning, mais la faible profondeur des données d'entraînement (7 ans) pourrait expliquer pourquoi le modèle n'a pas pleinement capté l'importance de l'ancienneté du contrat ou de l'âge de l'assuré par rapport à d'autres variables explicatives.

Les comparaisons entre la méthode machine learning et la méthode traditionnelle montrent que :

- La méthode machine learning évalue à la fois le risque de résiliation structurel (ancienneté de la police, âge, genre, etc.) et conjoncturel (fluctuations économiques et financières), souvent absent dans les lois de résiliation des assureurs pour les contrats UC car pénalisant dans les calculs d'engagements et non imposé par le régulateur.
- Utiliser un jeu de données historiques plus profond, incluant des souscriptions antérieures à 2000, permettrait une modélisation plus précise des algorithmes de machine learning, enrichissant la compréhension des résiliations structurelles du portefeuille.

Après avoir estimé l'écoulement du portefeuille de contrats dû aux résiliations, nous procédons au calcul de la valeur économique des flux futurs (VIF) en appliquant les taux de marge nette par contrat restant dans le portefeuille projeté. A partir de la capacité du modèle de machine learning à prédire les résiliations en fonction des fluctuations des indicateurs économiques et financiers, nous pourrions procéder à une valorisation selon les principes de la MCEV (Market Consistent Embedded Value) en stochastique dans un environnement risque neutre pour déterminer la valeur intrinsèque du portefeuille. Disposant d'un scénario déterministe d'environnement économique, nous procédons à la réalisation d'une valorisation selon les principes de la TEV (Traditional Embedded Value) prévoyant une marge pour risque dans le taux d'actualisation en supposant des hypothèses diverses (marge financière, coûts de gestion, besoin en capital et son coût d'immobilisation, etc.) sur le portefeuille étudié.

Nous avons synthétisé l'ensemble des projections dans un tableau :

En €		TEV (méthode traditionnelle)	TEV (XGBoost)
ANR		293 741	
VIF	**PVFP**	154 922	168 724
	CoC	-84 679	-92 223
Valorisation		**363 984**	**370 243**

Dans nos cas d'étude, la valeur d'un portefeuille de 9,8M€ de contrats UC à fin 2008 est estimée à 360-370K€ selon la méthode d'écoulement utilisée. La différence entre les deux méthodes n'est que de 2%, du fait d'un biais dans la modélisation de l'écoulement des contrats mentionné ci-avant. La faible profondeur des données empêche le modèle de machine learning de capter pleinement le risque de résiliation structurel, rendant ses estimations de résiliations légèrement inférieures à celles basées sur les taux historiques. La méthode traditionnelle ne capte quant à elle pas l'impact du pic inflationniste entre 2021 et 2023 contrairement au modèle de machine learning.

Cette approche montre la contribution du machine learning dans des contextes signifiants en s'adaptant à l'information disponible, offrant des outils complémentaires pour les décisions stratégiques en contexte transactionnel.

Cette démarche vise également à élargir les applications dans le contexte d'IFRS 17, qui introduit des principes de pilotage et de valorisation davantage fondés sur les caractéristiques spécifiques des assureurs liées à leur offre de produits et segmentation de clientèle, à condition que les hypothèses soient justifiées de manière crédible du fait de ce secteur régulé.

Elle invite également à une réflexion plus large sur l'utilisation des technologies nouvelles et innovantes dans l'industrie de l'assurance, incitant chercheurs et professionnels à explorer de nouveaux horizons, remettre en question les pratiques traditionnelles et envisager des approches novatrices pour répondre aux enjeux actuels et futur du secteur.

Synthesis Note

In France, life insurance policies in Unit-Linked (UL) accounts attract many investors due to their high potential returns and flexibility, despite the lack of capital guarantees. This flexibility allows policyholders to react quickly to market changes by redeeming or terminating their policies at any time.

However, this poses a challenge for insurers, requiring prudent and dynamic management of UL funds to maintain financial stability and profitability. Policyholder decisions to redeem or terminate can significantly impact the assets under management and management fee income. Insurers may also need to liquidate assets at lower prices to meet these demands.

To mitigate these risks, they conduct detailed behavioural analyses and actuarially calibrate policy termination laws to anticipate terminations. This allows them to adjust asset management, maintain stable profitability, and effectively meet policyholders' liquidity demands.

In insurance company transactions, the intrinsic value of the portfolio plays a crucial role in determining the target's value, significantly influencing the price the buyer is willing to pay. This value is directly affected by the probability of policy terminations. A relatively stable portfolio with a low termination rate is more valuable as it reflects more sustainable future income, which leads to a more secured acquisition. Conversely, a high termination rate indicates greater uncertainty about the durability of future income flows, making the portfolio less attractive.

During a transaction process, the information provided to potential buyers is often limited, and termination laws are generally not disclosed during the preliminary study phase. Typically, the potential buyer uses its own termination laws on the target portfolio, adapting them as much as possible for its valuation exercise. This creates a bias because the buyer's termination law often do not match with the target portfolio's clientele and may lack of conjunctural component, as it is not required by the regulator and penalizes insurer commitment calculations.

In this context, the main objective of the thesis is to provide an alternative approach for better calibrating the termination risk of a UL contract portfolio in a transactional context. The study uses machine learning methods to analyse termination data and develop a predictive model specific to the portfolio under study, considering the behavioural particularities of the clientele. It aims to propose a termination law calibration method more representative than the potential buyer's internal laws, often applied by default due to information restrictions during the due diligence and valuation phase before submitting the final offer to the sellers.

This tailored calibration proves strategic as it would allow a finer evaluation of the target portfolio through a better estimation of termination risk and could offer a significant competitive advantage in the transaction process.

Due to the strategic importance of these predictions and the financial stakes for potential buyers, the thesis also acknowledges the limitations of this approach in a transactional context. The limitations stem from the quality of available information and the difficulties of applying machine learning in a concrete transactional context. Indeed, in these situations, operational efficiency and execution speed are crucial due to tight study deadlines. Additionally, model interpretability and approval by the actuarial function are essential given the financial stakes.

The study focuses on analysing termination data and developing a predictive model that aims not only to provide comfort in a valuation exercise in a transactional context but also to propose an alternative method to traditional termination law calibration methods to improve the perception of termination risk in UL life insurance portfolios, considering clients' behavioural specifics in response to economic and financial fluctuations.

This dual perspective paves the way for new practices by increasingly integrating machine learning in the insurance sector, where data mastery is at the heart of the business.

Data analysis

The study uses the 'eusavingULnoPSperYr' dataset from the 'CASdatasets' package by Arthur Charpentier and Christophe Dutang. This dataset compiles information on Unit-Linked (UL) savings products, specifically non-profit sharing contracts from anonymous EU insurers. It is considered representative of a standard UL life insurance portfolio. The dataset includes data from 1999 to 2007, covering several significant phases of the European economic cycle, marked by periods of growth and recession, providing insights into subscriber trends and behaviours over nearly a decade. It is important to note that the dataset only includes entries related to new contracts and terminations between 1999 and 2007.

The objective is to analyse this dataset to identify subscriber behavioural trends in response to variations in economic and financial indicators, such as changes in the Consumer Price Index (CPI) or the European stock market index over different periods.

Data analysis is a crucial step in the approach, aiming to visually understand the factors influencing termination decisions and identify anomalies that could bias the models.

The initial phase of Exploratory Data Analysis (EDA) begins with visual analysis through various charts and visualizations, essential for spotting trends and anomalies and identifying key attributes for the future predictive model. This approach aims to capture critical factors in the termination process.

Next, we conduct a technical study to deepen the examination of the dataset's characteristics by analysing distributions, correlations, and other descriptive statistics to understand the interactions between variables. The EDA, including handling missing values and identifying outliers, will prepare the data for modelling, ensuring the model's reliability and efficiency.

The goal is to structure the dataset to optimize the predictive model's performance by carefully selecting variables and refining data preparation to improve prediction accuracy.

Building a predictive model

Developing a predictive model involves several key phases, from data preprocessing to model evaluation.

The first step, data preprocessing, transforms raw data into a format suitable for analysis, involving cleaning, normalization, and encoding to optimize input for machine learning models.

Next, the strategic selection of algorithms allows us to evaluate various models to find the most effective tool for our prediction. These include Logistic Regression for binary classification, Random Forests for their ability to handle complex relationships between a large number of explanatory variables, XGBoost (eXtreme Gradient Boosting) for its performance and ability to learn from its errors, and Neural Networks for their capacity to capture non-linear relationships.

Each model is evaluated based on several performance criteria (precision, recall, ROC curve, confusion matrix, interpretability).

Following this evaluation, an optimization phase is undertaken to fine-tune parameters and improve performance, ultimately selecting the XGBoost model for its ability to provide accurate and reliable predictions, with a better balance between various performance indicators.

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.6244	0.25	0.002	0.0039	0.6787
Desicion Tree	0.7706	0.3545	**0.3797**	0.3667	0.6166
Random Forest	0.8281	0.5136	0.3209	**0.395**	0.7643
XGBoost	**0.8426**	**0.6073**	0.283	**0.3861**	**0.7873**
MLP neural networks	**0.8414**	**0.6076**	0.2621	0.3662	**0.7772**

Below the key reasons for recommending the XGBoost model in our study case:

- **F1 Score:** XGBoost model showed a high F1 score, indicating an excellent balance between precision and recall. This means the model can correctly predict a large number of positive cases while limiting false positives.
- **Precision and Recall:** XGBoost model demonstrated high precision and recall, crucial for applications where minimizing classification errors is important, such as predicting termination risk.
- **AUC Score:** With a high AUC score, XGBoost model proved its ability to effectively distinguish between positive and negative classes. A high AUC indicates the model's robustness across different classification thresholds.
- **Confusion Matrix:** Confusion matrices revealed that XGBoost model has the best balance between true positives (TP) and true negatives (TN), with a reduced number of false positives (FP) and false negatives (FN). This indicates an excellent ability to correctly classify instances.
- **ROC Curve**: XGBoost's ROC curve, closer to the upper left corner, illustrates its superiority in terms of true positive rate versus false positive rate. This demonstrates its ability to maximize true positives while minimizing false positives.

XGBoost model stands out for its overall superior performance across all evaluated criteria. Beyond structural explanatory variables (policyholder age and contract duration), the analysis of the most influential explanatory variables in the XGBoost model highlights two key variables:

- **CPI.relvar2yr:** Indicates the inflation evolution over two years through the CPI. CPI fluctuations affect purchasing power, influencing consumers' ability or willingness to maintain premium payments, directly impacting termination risk.
- **EUidx.relvar2yr:** Represents the long-term evolution of the European stock market index over two years, indicating the impact of economic trends on termination decisions. Significant variations can reveal adjustments in the economic environment, influencing policyholders' risk perception and financial stability.

Projecting termination behaviour and portfolio valuation perspectives

Our approach aims to estimate the future evolution of terminations within the UL life insurance portfolio to evaluate the progressive degradation of the portfolio, i.e., the reduction in the number of active contracts without considering new contracts. We assume no additional contributions or new contracts in our projections to make the analysis. As the dataset presents historical data from 1999 to 2007, we will focus on active contracts at the end of this period as the basis for our valuation.

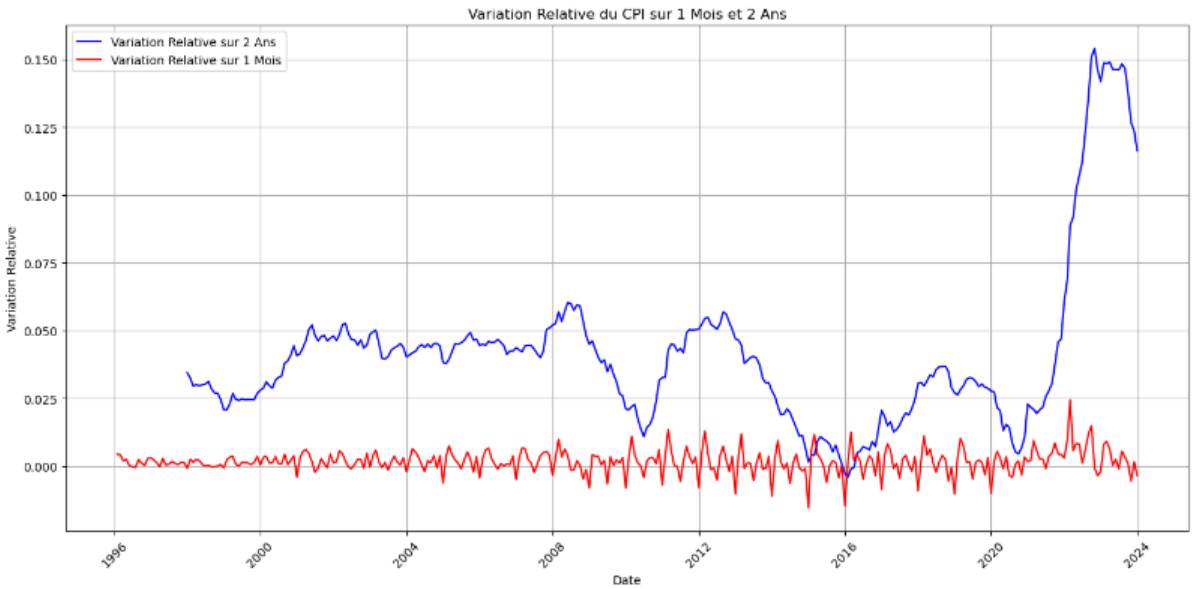
The next step involves feeding the initially trained model on historical data (here XGBoost) with one or more scenarios of the economic environment over a relevant projection horizon to establish the portfolio's runoff. We project the most influential economic and financial variables of our model, namely

the CPI and the European stock market index (EUidx), as well as specific policyholder characteristics, to predict their impact on termination decisions.

We propose several approaches for projecting these economic and financial variables: a deterministic scenario, time series to capture economic cycles, or stochastic methods. Insurers generally use Economic Scenario Generators, which are recommended for creating stochastic projections of explanatory variables to serve as input data for the predictive model.

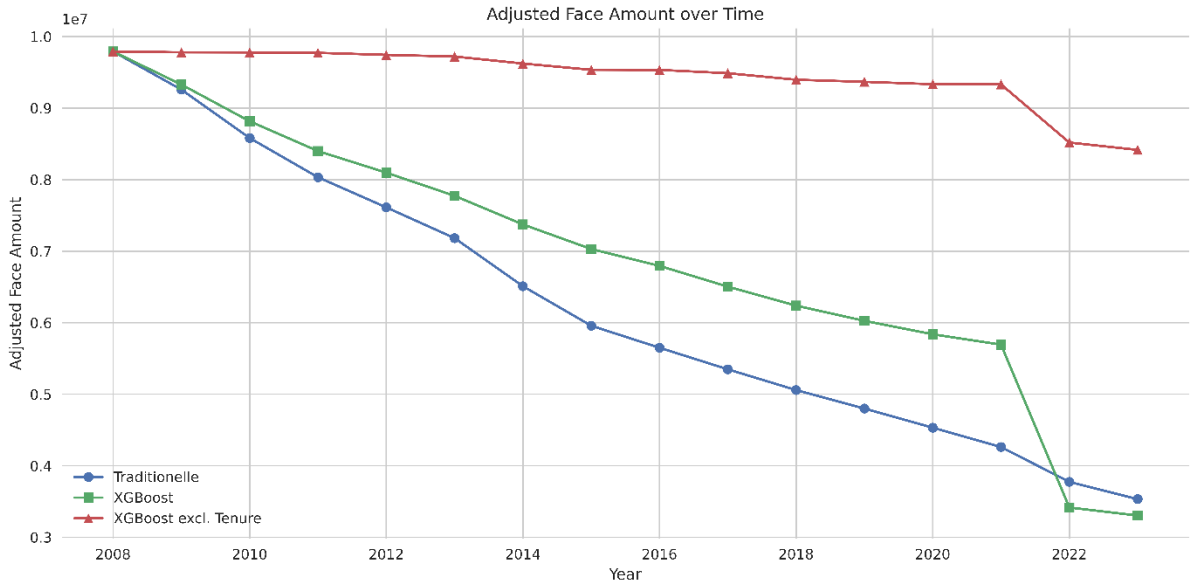
The available data constrain us to value this portfolio as of the end of 2007. Instead of generating projections for the explanatory variables, we opted for a simplified and more efficient approach using actual economic and financial data between 2008 and 2023. Note that this approach can be followed up to the desired valuation date as long as data is available up to that date. We then use the Economic Scenario Generator to project the explanatory variables as input data for the model. In our case, we rely on the historical developments of the Euro Stoxx 50 index, representing the performance of major companies in the euro area, and the CPI, which measures inflation.

Valuing a UL life insurance portfolio typically requires a projection horizon of at least 40 years. For simplicity and efficiency, we limit ourselves to a 15-year period, using only actual data for our variables, covering at least one economic cycle and revealing the impacts of Covid. This period is sufficiently long to cover several economic times to make an accurate analysis of trends, although it remains short to avoid uncertainty and too much complexity.



Relative change over 2 years (blue) and 1 month (red) in the CPI (Consumer Price Index)

The trained XGBoost model is then used to estimate future terminations based on a deterministic scenario of explanatory variables (CPI and Euro Stoxx 50) over the 2008-2023 horizon. For comparison purposes, we simulated the run-off of the contract portfolio using the traditional method of simply using the portfolio's historical termination rates.



Run-off Portfolio erosion - Traditional method vs. XGBoost Machine Learning method (and for illustrative purposes XGBoost excl. Tenure)

The comparison of portefolio erosions shows that the traditional method, based on the intrinsic evolution of portfolio characteristics (contract duration and policyholder age), presents a stable and nearly linear erosion. In contrast, the XGBoost method reveals a significant lapse in contracts in response to the inflation peak, a factor ignored by the traditional method. The latter does not capture the impact of the inflation peak on policyholder termination behavior.

Ultimately, the results are similar: the traditional method predicts 64% terminations over 15 years, while XGBoost model estimates 66%, with 14% attributed to the inflation peak during Covid. Without this effect, XGBoost model would predict approximately 50-55% terminations. The explanation for this discrepancy is complex due to the "black box" nature of machine learning models, but the limited depth of the training data (7 years) might explain why the model did not fully capture the importance of contract duration or policyholder age compared to other explanatory variables.

Comparisons between the machine learning method and the traditional method show that:

- The machine learning method evaluates both structural termination risks (policy duration, age, gender, etc.) and conjunctural risks (economic and financial fluctuations), often absent in insurers' termination laws for UL contracts as they are penalizing in liability calculations and not required by the regulator.
- Using a deeper historical dataset, including subscriptions prior to 2000, would allow for more precise machine learning algorithm modelling, enriching the understanding of the portfolio's structural terminations.

After estimating the portfolio's termination flow, we calculate the economic value of future cash flows (VIF) by applying the net margin rates per remaining contract in the projected portfolio. Based on the learning model's ability to predict terminations according to economic and financial indicator fluctuations, we could proceed with a valuation according to the principles of MCEV (Market Consistent Embedded Value) in a stochastic, risk-neutral environment to determine the portfolio's intrinsic value. With a deterministic economic environment scenario, we conduct a valuation following TEV (Traditional Embedded Value) principles, incorporating a risk margin in the discount rate by assuming various hypotheses (financial margin, management costs, capital requirements, and their immobilization costs, etc.) on the studied portfolio.

We have summarized all projections in a table:

En €		TEV (méthode traditionnelle)	TEV (XGBoost)
ANR		293 741	
VIF	**PVFP**	154 922	168 724
	CoC	-84 679	-92 223
Valorisation		**363 984**	**370 243**

In our case studies, the value of a €9.8M UL contract portfolio at the end of 2008 is estimated at €360-370K, depending on the outflow method used. The difference between the two methods is only 2%, due to the aforementioned bias in the outflow modelling of contracts. The limited depth of the data prevents the machine learning model from fully capturing the structural termination risk, resulting in slightly lower termination estimates compared to those based on historical rates. The traditional method, on the other hand, fails to capture impact of the inflation peak between 2021 and 2023, unlike the machine learning model.

This approach demonstrates the contribution of machine learning in contexts that can make sense by adapting to the available information, offering complementary tools for strategic decision-making in transactional contexts.

This method also aims to broaden applications in the context of IFRS 17, which introduces management and valuation principles more based on the specific characteristics of insurers related to their product offerings and customer segmentation, provided that the assumptions are credibly justified in this regulated sector.

It also invites a broader reflection on the use of new and innovative technologies in the insurance industry, encouraging researchers and professionals to explore new horizons, challenge traditional practices, and consider innovative approaches to address the current and future challenges of the sector.

Remerciements

Tout d'abord, je tiens à remercier Yann CHOQUET, mon tuteur en entreprise, pour son accompagnement, ses conseils, ses relectures et pour l'ensemble des échanges menés pour ce mémoire. Plus généralement, je tiens à le remercier pour son implication et son aide. Ces remerciements s'appliquent également à Maryline MADALENO notamment pour toutes les pistes de réflexion apportées lors de nos échanges.

J'aimerai ensuite remercier Georges Louis GONCALVES, mon tuteur académique, pour son encadrement, ses relectures et ses conseils avisés. Je remercie également le corps professoral de l'IRM et de la Sorbonne Université pour la qualité de l'enseignement dispensé.

Je suis particulièrement reconnaissant à l'équipe M&A du Crédit Mutuel Arkéa pour son accueil et son accompagnement. Je remercie également à tous les membres de l'équipe pour leur soutien qui a contribué au bon déroulement de ce travail.

Enfin, j'ai une pensée particulière pour mes proches pour leur soutien constant tout au long de mes études. Je tiens également à remercier ma conjointe pour son soutien inestimable et constant.

Table des matières

Introduction

Contexte et importance de l'étude

L'assurance vie en Unités de Compte (UC) se distingue comme pilier majeur du marché financier français, attirant un large éventail d'investisseurs, du grand public aux clients patrimoniaux. Ce type de produit d'assurance vie se démarque par sa capacité à offrir un accès à des supports d'investissement dynamiques, sans pour autant garantir le capital investi, offrant un potentiel de rendement attractif malgré un risque plus élevé.

En France, les contrats d'assurance vie sont fiscalement avantageux, ce qui en fait un investissement prisé. Les Unités de Compte (UC) se distinguent par leur liquidité, offrant aux souscripteurs une flexibilité pour réagir aux fluctuations du marché et ajuster leur stratégie. Cette liquidité, garantie par l'assureur, rend les UC attrayantes pour ceux recherchant performance et accessibilité de leurs actifs.

Cependant, cette flexibilité s'accompagne d'un défi pour les assureurs : la compréhension du risque de résiliation des contrats. La possibilité pour les souscripteurs de racheter ou de résilier leurs contrats à tout moment impose aux compagnies d'assurance une gestion prudente et réactive de leurs investissements au titre de ces supports en UC. La résiliation anticipée des contrats peut en effet (i) impacter significativement la stabilité financière et la rentabilité des assureurs qui se rémunèrent sur le niveau d'encours des assurés et (ii) poser un problème de liquidité lorsqu'il y a un déséquilibre de l'offre et la demande sur un actif en particulier, comme lors de la baisse des parts de Sociétés Civiles de Placement Immobilier à la fin de l'année 2023.

Ce sujet étant crucial pour chaque assureur, ces derniers réalisent de nombreuses analyses afin de calibrer le plus finement possible les lois de résiliations pour leur portefeuille. La connaissance de leur clientèle couplée à la profondeur importante des données historiques de clientèle détenues par les assureurs leur permet d'ajuster leur loi de résiliation année après année.

Dans un contexte transactionnel, la valeur intrinsèque du portefeuille d'une compagnie d'assurance est une composante prépondérante de la valorisation de la cible assurantielle, celle-ci définissant en partie le prix proposé par le potentiel repreneur. La valeur que le repreneur est prêt à investir pour acquérir ce portefeuille est en partie liée à la durée des contrats en portefeuille. De ce fait, la valeur intrinsèque du portefeuille, souvent désignée par le terme "Value In Force", est directement impactée par la probabilité de résiliation des contrats. Un portefeuille caractérisé par une forte stabilité clientèle et un faible taux de résiliation est généralement plus valorisé, reflétant une source de revenus future plus sûre. À l'inverse, un portefeuille avec un taux de résiliation élevé peut être considéré comme moins attrayant, car il suggère une volatilité plus importante et une incertitude quant aux flux de revenus futurs.

Lorsqu'une compagnie d'assurance envisage la vente de son portefeuille, ou lorsqu'elle fait elle-même l'objet d'une acquisition, l'information mise à disposition des potentiels repreneurs se révèle être souvent limitée et les lois de résiliations sur le portefeuille cible ne sont en général pas communiquées durant la phase d'étude en amont de la transaction. Dans les faits, il est usuel que le potentiel repreneur utilise ses propres lois de résiliation qu'il applique au portefeuille cible, en l'adaptant dans la mesure du possible, pour son exercice de valorisation. Cela génère un biais car il est hautement probable que la loi de résiliation du repreneur ne soit pas représentative de la clientèle du portefeuille cible. De plus, pour les contrats en UC, les lois de résiliation des assureurs peuvent ne pas inclure de composante conjoncturelle, mais uniquement une composante structurelle, car non imposée par le régulateur et pénalisante dans les calculs d'engagement.

Dans ce cadre, l'amélioration de la prédiction du risque de résiliation des contrats spécifique au portefeuille de la cible serait un atout stratégique majeur pour les repreneurs. Un modèle prédictif, qui a été entrainé sur les données historiques du portefeuille de la cible en question, pourrait fournir des indications précieuses sur la valeur du portefeuille, permettant ainsi aux assureurs de maximiser la valeur de leur offre dans le cadre de transactions. Une évaluation plus fine du portefeuille cible pourrait ainsi offrir un avantage concurrentiel stratégique notoire dans le processus de transaction qui est souvent compétitif à l'achat.

Notre étude se concentre sur l'analyse des données de résiliation et le développement d'un modèle prédictif qui vise non seulement à donner du confort sur un exercice de valorisation dans un contexte transactionnel, mais aussi à proposer une démarche alternative aux méthodes traditionnelles de calibrage des lois de résiliation afin d'améliorer la perception du risque de résiliation des portefeuilles d'assurance vie en UC prenant en compte les spécificités comportementales des clients.

Cette double perspective ouvre la voie à de nouvelles pratiques par l'intégration grandissante du machine learning dans le secteur des assurances, où la maîtrise de la donnée est au cœur de l'activité.

Dans le cadre de cette étude, nous proposons d'explorer l'application des techniques de machine learning pour analyser les données des contrats d'assurance vie en UC, en nous focalisant sur des données publiques d'un assureur européen entre 1999 et 2007. L'objectif est triple : identifier les tendances et les comportements de résiliation des assurés, développer un modèle prédictif capable d'anticiper ces choix et l'appliquer afin d'estimer les résiliations pour en déduire la valeur du portefeuille dans un cadre transactionnel.

Cette approche vise à doter les assureurs d'outils analytiques stratégiques pour l'estimation du risque de résiliation dans le cadre de la valorisation des engagements dans un contexte transactionnel, mais aussi dans d'autres cadres comme IFRS 17 qui introduit notamment des principes de pilotage et de valorisation qui s'appuient davantage sur la base du comportement des assurés.

I. Objectifs et portée de l'étude

1. Démarche de l'étude

Ce mémoire propose une démarche qui permettrait d'apporter une contribution supplémentaire à l'aide du machine learning dans la maîtrise du risque de résiliation dans les assurances vie en Unités de Compte (UC) en nous focalisant ici sur un contexte transactionnel. L'objectif est de fournir une méthode alternative de calibrage des lois de résiliation de contrat grâce aux méthodes de machine learning, dont l'usage s'avère pertinent dans un contexte transactionnel.

En effet, l'information communiquée par les vendeurs sur la cible est usuellement très limitée et les lois de résiliation ne sont en général pas transmises lors du processus de due diligences et de valorisation en amont de la transaction. Il est ainsi usuel que le potentiel acheteur utilise ses propres lois de résiliation pour estimer l'écoulement de contrats du portefeuille cible. Cependant, cela génère un biais car les lois de résiliation du potentiel acheteur ne sont en général pas représentatives de la clientèle du portefeuille cible. De plus, concernant les contrats en UC, les lois de résiliation des assureurs peuvent ne pas comporter de composante de résiliation conjoncturelle, en plus de la composante de résiliation structurelle, car à la fois non imposée par le régulateur et pénalisante dans les calculs prudentiels. Dans un contexte transactionnel, cela est bien pris en compte afin de converger vers une vision réaliste qui comprend toutes les composantes du risque de résiliation. Il est ainsi primordial d'estimer avec précision le risque de résiliation.

Dans ce contexte, l'analyse des résiliations repose sur l'exploitation des données historiques, dont la granularité et les critères varient selon les compagnies d'assurance, influençant ainsi la pertinence des modèles prédictifs. Parmi les approches possibles, la régression logistique constitue une méthode classique pour analyser ces comportements. Il est d'ailleurs possible d'intégrer les effets conjoncturels dans une régression logistique en ajoutant des indicatrices (variables binaires ou catégoriques). Cependant, les algorithmes de machine learning, tels que XGBoost ou les Forêts Aléatoires, offrent une capacité intéressante à intégrer des facteurs complexes, notamment l'évolution des indicateurs économiques et les interactions non linéaires entre variables. Ces modèles permettent ainsi d'améliorer la précision des prévisions en capturant les spécificités comportementales des assurés du portefeuille étudié, en tenant compte de multiples paramètres avec leurs potentielles interdépendances, tels que l'âge, la durée du contrat, le niveau d'exposition, ainsi que les dynamiques économiques, financières et sociétales.

Les modèles de machine learning utilisés pour prédire les résiliations vont ainsi capturer à la fois les résiliations structurelles, selon les moments de vie des assurés / ancienneté des contrats / avantages

fiscaux / etc., et les résiliations conjoncturelles relatives aux fluctuations de marché et d'indicateurs économiques par rapport à leur niveau d'exposition.

Dans un contexte transactionnel, l'accès à des données détaillées sur le portefeuille de la cible reste une condition cruciale pour assurer un calibrage précis des lois de résiliation. Par ailleurs, en raison des contraintes contextuelles de temps d'étude très limité lié au contexte transactionnel, l'application des lois de résiliation des acquéreurs offre souvent une vision peu représentative du portefeuille cible. Les algorithmes de machine learning semblent permettre une approche plus adaptée dans ce type de contexte. En effet, ces méthodes permettent de traiter un volume important de données tout en s'adaptant aux délais restreints imposés par les transactions. Contrairement aux approches classiques, qui nécessitent une analyse approfondie préalable des bases de données avant tout calibrage du risque de résiliation, le machine learning permet d'optimiser l'exploitation des données dans un temps limité, répondant ainsi aux exigences des processus transactionnels.

Dans le cadre de la démarche proposée dans cette étude, nous commencerons par analyser les données disponibles, qui pourraient s'apparenter aux données communiquées par les vendeurs relatives au portefeuille de contrats de la cible.

Après avoir réalisé une analyse exploratoire de ces données afin d'en tirer des premières observations, nous procéderons à la préparation de cette base de données pour permettre un meilleur apprentissage des modèles.

Une fois les modèles de machine learning entrainés sur la base de données historiques du portefeuille, nous procéderons à la sélection du modèle de prédiction le plus pertinent par comparaison des performances de chacun et nous visualiserons les variables explicatives les plus importantes qui semblent avoir un rôle prépondérant dans le choix des assurés quant à leur décision de résiliation de contrat.

Ensuite, nous proposerons des méthodes pour projeter ces variables explicatives sur un horizon de valorisation pertinent pour l'évaluation du portefeuille étudié. L'application du modèle prédictif de résiliation sur la base de ces projections de variables explicatives permettra ainsi de prédire l'écoulement du portefeuille, selon l'évolution des critères inhérents aux assurés et selon les fluctuations des indicateurs économiques et financiers.

Nous proposons de manière subséquente la valorisation de ce portefeuille grâce à la trajectoire d'écoulement prédite par le modèle de machine learning, en appliquant un niveau de chargement sur l'encours afin d'évaluer la marge future intrinsèque du portefeuille.

Il est par ailleurs important de mentionner que cette étude vise uniquement à proposer une démarche permettant d'apporter une vision alternative aux méthodes traditionnelles du calibrage du risque de résiliation et que celle-ci comporte des limites à la fois au regard de la qualité de l'information disponible

mais aussi dans la mesure où les méthodes de machine learning sont à l'heure actuelle encore peu déployées chez les assureurs du fait notamment de multiples contraintes diverses, telles que l'interprétabilité des modèles et des résultats mais aussi l'inscription dans la réalité de ce qui serait implémenté et validé par la fonction actuarielle.

A. Les données disponibles pour notre étude :

Les données qui serviront de base pour notre étude proviennent du dataset 'eusavingULnoPSperYr', issu du package 'CASdatasets' compilé par Arthur Charpentier et Christophe Dutang. Ce dataset contient des informations spécifiques par client de collectes sur des produits d'épargne Unit-Linked purs (uniquement des UC), sans participation aux bénéfices, provenant d'assureurs vie anonymes de l'Union Européenne, sur la période allant de 1999 à 2007.

Ce dataset est considéré comme représentatif d'un portefeuille typique d'une compagnie d'assurance vie en UC. Il s'agit d'un portefeuille en cours de constitution au regard de la profondeur du dataset relativement limitée (8 années).

Nous avons donc à notre disposition un portefeuille de contrats UC à fin 2007, issu de collectes sur la période allant de 1999 à 2007.

L'idée de l'étude est de proposer une démarche alternative pour l'estimation du risque de résiliation structurel et conjoncturel de ce portefeuille à fin 2007, en utilisant les méthodes de machine learning, dans le but de valoriser ce portefeuille à fin 2007. Les données que nous avons nous contraignent à réaliser la valorisation de ce portefeuille à fin 2007. A noter que cette démarche peut être suivie à la date de valorisation souhaitée, sous réserve de disposer des données jusqu'à cette date de valorisation. Il convient ensuite d'utiliser un Générateur de Scénarios Economiques pour projeter les variables explicatives comme données d'entrée du modèle.

B. L'analyse descriptive des données :

L'analyse approfondie du dataset représente une étape fondamentale de cette étude.

Cette phase vise à déployer une approche analytique détaillée, permettant de dégager une compréhension initiale des facteurs influençant la résiliation des contrats d'assurance vie en UC. Cette analyse s'appuie sur une Exploratory Data Analysis (EDA) et représente une étape préalable au développement ultérieur d'un modèle prédictif fiable.

La première étape de cette analyse consistera en une exploration visuelle des données. L'utilisation de graphiques, de diagrammes et de visualisations interactives permettra d'identifier rapidement les tendances, les anomalies et les modèles potentiels dans les données. Cette approche visuelle est essentielle pour comprendre les facteurs qui jouent un rôle majeur dans la résiliation des contrats. Elle aidera à déterminer quels sont les attributs les plus pertinents à inclure dans le modèle prédictif.

Au-delà de l'exploration visuelle, une analyse plus technique et détaillée des différents champs du dataset sera entreprise. Cette partie de l'analyse se concentrera sur l'examen des caractéristiques individuelles des données, telles que les distributions, les moyennes, les écarts-types, et d'autres statistiques descriptives. Une attention particulière sera accordée à l'identification et à l'analyse des corrélations entre les différentes variables. Comprendre comment ces variables interagissent entre elles est crucial pour construire un modèle prédictif qui capture fidèlement la complexité des comportements de résiliation.

L'EDA jouera un rôle clé dans cette phase. Elle impliquera une investigation des données pour en extraire des déductions sur le comportement des assurés. Cette analyse exploratoire comprendra la détection et le traitement des valeurs manquantes, l'identification des valeurs aberrantes, et l'analyse des distributions de variables. Une analyse multivariée permet notamment d'observer les éventuelles corrélations entre les variables explicatives, qu'il convient d'ajuster si besoin. L'objectif est de préparer les données de la manière la plus complète et la plus propre possible, afin de minimiser les biais et d'optimiser l'efficacité du modèle prédictif.

Enfin, cette phase d'analyse des données aboutira à la préparation des données pour leur utilisation dans le modèle prédictif. Cela impliquera des décisions cruciales sur la sélection des variables, la transformation des données, et la création de nouvelles variables si nécessaire. L'objectif est de s'assurer que le dataset est structuré de manière à maximiser la performance du modèle, en réduisant au maximum les biais et en augmentant la précision des prédictions.

C. Développement du modèle prédictif de résiliation :

Le développement d'un modèle prédictif de résiliation est une étape clé de cette étude, visant à fournir un outil analytique capable d'estimer les comportements de résiliation des contrats d'assurance vie en UC.

Cette phase implique plusieurs étapes critiques, allant du prétraitement des données à l'entraînement et au test du modèle, en passant par la sélection des algorithmes les plus adaptés.

Avant même de commencer à construire le modèle, une étape essentielle de data preprocessing est nécessaire. Cette phase vise à transformer les données brutes en un format adapté pour l'analyse et la modélisation. Le preprocessing inclura le nettoyage des données et le formatting des données pour les rendre les plus exploitables pour les modèles de machine learning. Il comprendra également la normalisation ou la standardisation des variables pour assurer que le modèle ne soit pas biaisé par des échelles de mesure différentes. De plus, des techniques telles que l'encodage des variables catégorielles et la sélection des caractéristiques seront appliquées pour optimiser la qualité des données entrantes dans le modèle.

Le choix des algorithmes de machine learning est crucial. Dans le cadre du développement de notre modèle prédictif de résiliation, une approche comparative sera adoptée. Plutôt que de se limiter à un seul type de modèle de machine learning, plusieurs modèles seront testés et évalués pour déterminer lequel est le plus adapté à notre cas d'étude. Cette stratégie permet de garantir que le modèle choisi offre la meilleure combinaison possible de précision et d'efficacité pour prédire les comportements de résiliation dans les assurances vie épargne en UC. En effet, selon les cas, certains modèles peuvent être plus performants que d'autres.

Parmi les modèles envisagés, on peut citer la Régression Logistique, les Forêts Aléatoires, XGBoost (eXtreme Gradient Boosting), les Réseaux de Neurones / Deep Learning. XGBoost est réputé pour sa vitesse et sa performance, en particulier dans les compétitions de data science. Cependant, dans notre cas, les Réseaux de Neurones peuvent aussi être performants étant donné qu'ils parviennent à capturer des relations non linéaires complexes, bien que souvent plus exigeants en termes de données. Le modèle de Forêts Aléatoires est souvent aussi très performant dans les problématiques de classification. La Régression Logistique peut s'avérer pertinente dans les cas où la variable à prédire suit une logique binaire comme ici (résiliation ou non).

Chaque modèle sera évalué en fonction de plusieurs critères de performance clés (Précision, Rappel, Aire sous la courbe ROC, Matrice de confusion, Interprétabilité).

Après une évaluation initiale des performances des modèles, les modèles seront soumis à une phase d'optimisation pour affiner leurs paramètres et améliorer leur performance. Cette étape impliquera des ajustements fins, tels que la modification des paramètres de régularisation, le réglage des taux d'apprentissage, ou la sélection des caractéristiques.

Le modèle final sera choisi non seulement sur la base de sa performance statistique, mais aussi en considérant sa pertinence pratique pour le cas d'étude, y compris la facilité de mise en œuvre et l'interprétabilité des résultats.

D. Projection des comportements de résiliation et valorisation du portefeuille :

La prochaine étape consiste à projeter les résiliations afin d'obtenir une vision prospective de l'écoulement des contrats en portefeuille, dans le but d'aboutir à la valorisation du portefeuille cible.

Le modèle prédictif, une fois entraîné et validé, sera appliqué sur la base de scénarios d'évolution des variables explicatives les plus importantes. Nous envisagerons différentes méthodes pour prédire l'évolution de ces caractéristiques. Compte tenu de leur nature temporelle, des modèles spécifiques aux séries temporelles, comme SARIMAX, pourraient être utilisés pour capturer les tendances et interactions entre variables économiques (ex. inflation, indices boursiers, taux d'intérêt). Ces approches permettent d'intégrer des facteurs exogènes influençant directement l'évolution des variables explicatives. En complément, une approche stochastique via un Générateur de Scénarios Économiques

(GSE) est souvent privilégiée par les assureurs pour simuler des conditions économiques futures et modéliser l'incertitude. En combinant ces différentes approches, il devient possible d'obtenir des projections plus robustes et adaptées aux diverses conditions de marché. Ces scénarios sont conçus pour simuler diverses conditions économiques et autres facteurs déterminants qui pourraient influencer les décisions de résiliation des souscripteurs. En intégrant des variables telles que les fluctuations des marchés financiers, les changements de politique monétaire, ou même des événements socio-économiques majeurs, ces projections stochastiques permettraient de modéliser des environnements futurs possibles.

Dans le cadre de notre étude et en l'absence de Générateur de Scénarios Economiques, nous retiendrons un scénario déterministe sur 2008-2023 par souci de simplification et d'efficience pour la suite de l'étude. Dans les cas concrets transactionnels, la date de valorisation étant généralement la dernière date d'arrêté comptable, les assureurs peuvent utiliser un Générateur de Scénarios Economiques afin de projeter les variables explicatives sur un horizon de projection défini pour l'exercice de valorisation (en général sur entre 50 et 70 ans).

L'un des aspects innovants de cette approche dans le contexte transactionnel est sa capacité à adapter les projections de résiliation aux comportements spécifiques des assurés du portefeuille. Plutôt que de se baser sur des hypothèses historiques généralisées, le modèle prend en compte les tendances et comportements de la clientèle observés au sein du portefeuille spécifique. Cela permet d'obtenir des prévisions de résiliation qui reflètent plus fidèlement la manière dont les souscripteurs pourraient réagir face à différents scénarios économiques ou changements de marché. Dans un contexte transactionnel, cela évite l'application des règles de résiliation de l'acquéreur potentiel, du fait du manque de temps de traitement de l'information, pour calibrer les résiliations du portefeuille de la cible, ce qui pourrait introduire des biais en raison des différences comportementales des clientèles. A noter que cela implique l'accès à l'information détaillée historique, telles que les *client tapes* (bases de données clients) et les AuM tapes (bases des encours) dans un historique de profondeur suffisant, qui sont en général transmises lors des processus transactionnels et qui sont une source d'informations riches et détaillées sur les contrats des clients.

La projection de l'écoulement du portefeuille en vue d'une perspective de valorisation du portefeuille est une autre étape clé dans cette approche. Pour les besoins de notre étude et dans le cadre d'une valorisation selon une approche du calcul de la VIF (*value in force*), nous proposons ici de nous focaliser uniquement sur l'évolution du portefeuille sans affaires nouvelles (*backbook*). En utilisant les prévisions de résiliation issues du modèle, il est possible d'estimer l'écoulement prévisionnel du portefeuille selon les scénarios de projetés des variables explicatives. A partir des niveaux de chargement de l'assureur cible, il est ensuite possible de mesurer l'impact financier des résiliations futures sur la valeur du

portefeuille. Cette évaluation prend en compte les pertes directes dues aux résiliations de contrats, mais aussi les implications à long terme sur les flux de trésorerie et la rentabilité.

En fournissant une estimation de la valeur future du backbook par une compréhension de la dynamique de résiliation davantage représentative de la population du portefeuille étudié, les acquéreurs potentiels pourraient mieux appréhender la valeur intrinsèque de la cible et planifier leur stratégie financière par rapport à leur offre d'achat. Cette approche offre aux acteurs du milieu une perspective stratégique précieuse.

Pour challenger notre approche, nous avons comparé l'écoulement de contrats, et la valorisation qui en résulte, issus des méthodes de machine learning par rapport à ceux issus des méthodes traditionnelles, pour lesquelles nous avons considéré le taux de résiliation historique constaté sur ce portefeuille pour projeter l'écoulement et, in fine la valorisation du portefeuille.

E. Perspectives et applications

Le domaine de la prédiction de résiliation dans les assurances vie en UC est vaste et en constante évolution. Bien que la prédiction par les méthodes de Machine Learning est désormais devenue courante, l'objectif de cette étude est de contribuer à enrichir cette approche en apportant de nouvelles perspectives d'application dans lesquelles ces méthodes s'avèrent particulièrement intéressantes. Elle ouvre des perspectives nouvelles et stimulantes pour des applications futures.

Ces méthodes offrent la possibilité de modéliser des comportements complexes et de capter des tendances subtiles qui pourraient échapper aux approches traditionnelles dans certains cas bien spécifiques. En intégrant ces techniques dans l'analyse de données, les équipes pourraient obtenir des prévisions plus précises et adaptées aux spécificités de chaque portefeuille, ce qui peut se révéler être une information cruciale dans le cadre de travaux actuariels.

L'aptitude du machine learning à intégrer un grand nombre de variables explicatives, qui enrichit d'autant plus la compréhension des particularités du portefeuille, constitue également un atout majeur et prometteur de ces outils qui ouvre un large éventail de possibilités d'application. Chaque portefeuille a ses particularités, influencées par des facteurs tels que la démographie des souscripteurs, les choix d'investissement, et les politiques de gestion des risques. Les modèles de machine learning, avec leur capacité à traiter et à apprendre de grandes quantités de données, peuvent être ajustés pour refléter ces particularités, offrant ainsi une évaluation plus précise et personnalisée de la valeur et des marges futures issues de ce portefeuille.

Cette approche pourrait être optimisée afin d'améliorer sa facilité d'implémentation opérationnelle, notamment dans le cadre transactionnel souvent régi par des contraintes diverses de temps, de moyens ou de données disponibles.

Cette approche vise aussi à ouvrir les applications dans le cadre de IFRS17 qui introduit des principes de pilotage d'activité et de valorisation davantage fondés sur les spécificités des assureurs par rapport à leur activité, donc leur positionnement de produit, et de segment client, sous réserve de crédibilité dans la justification des hypothèses prises.

Elle invite également à une réflexion plus large sur l'application des technologies de pointe dans le secteur des assurances. Elle encourage les chercheurs et les professionnels à explorer de nouvelles voies, à questionner les méthodes établies, et à envisager des solutions innovantes pour les défis actuels et futurs du secteur.

2. Limites de l'étude

Dans le cadre de notre étude, malgré les dernières avancées en matière de machine learning et l'adoption grandissante de ce type d'outils, il est essentiel de reconnaître les limites inhérentes à l'application opérationnelle de la modélisation prédictive de résiliation de contrats d'assurance dans l'environnement règlementé dans lequel opèrent les assureurs, ainsi que les limites de l'utilisation de ces outils dans le contexte transactionnel compte tenu de contraintes spécifiques variées.

Cette prise de conscience nous permet d'aborder notre recherche avec une perspective conservatrice, en tenant compte à la fois des progrès technologiques et des défis pratiques.

A. Avancées récentes dans la modélisation prédictive :

Les dernières années ont vu des progrès significatifs dans le domaine de la modélisation prédictive, notamment grâce à l'évolution des techniques de machine learning et de l'analyse de données de plus en plus sophistiquées, offrant une meilleure capacité à analyser des ensembles de données complexes et à en extraire des prédictions.

Dans le secteur des assurances, ces avancées ont permis de mieux comprendre et anticiper les comportements de résiliation, en identifiant des motifs et les tendances comportementales qui étaient auparavant plus compliqués à saisir.

Cependant, ces innovations ne sont pas sans défis. L'un des principaux obstacles est la qualité et la disponibilité des données. Les modèles prédictifs sont fortement dépendants de la quantité et de la fiabilité des données historiques. Dans certains cas, les données peuvent être incomplètes, biaisées ou ne pas refléter fidèlement les conditions actuelles du marché, ce qui peut limiter la précision des prédictions.

Dans le contexte transactionnel, il est possible que cette donnée soit peu exploitable ou que celle-ci nécessite des traitements pour compiler une base de données solide et pertinente avant l'application des méthodes de machine learning. En général, les bases de données clients (*client tape*) des banques sont très détaillées et rassemblent une mine d'information très intéressante sur la clientèle pour ces études. Le niveau d'information relative au contrat d'assurance n'est cependant pas très riche et cela nécessite souvent de fusionner cette base de données de clients avec la base de données de l'assureur ou du distributeur (*AuM tape*) détaillant l'ensemble des données relatives aux contrats d'assurance vie en tant que telles.

B. Interprétabilité des modèles et équilibre pratique :

Un autre aspect crucial est l'interprétabilité des modèles. Alors que les techniques de machine learning deviennent de plus en plus complexes, il devient également plus difficile de comprendre comment ces modèles raisonnent et arrivent à leurs conclusions. Cette notion de "boîte noire" pose un problème, surtout dans un domaine réglementé comme l'assurance, où les décideurs doivent souvent justifier leurs actions et leurs décisions par des hypothèses et des méthodes interprétables.

L'importance de trouver un équilibre entre la précision technique des modèles et leur applicabilité pratique ne peut être sous-estimée. Un modèle extrêmement précis mais incompréhensible peut s'avérer moins utile concrètement qu'un modèle moins sophistiqué mais plus transparent et facile à interpréter.

La capacité d'un modèle à être intégré dans les processus décisionnels existants et à être utilisé efficacement par les professionnels de l'assurance est essentielle. Cela reflète en effet la capacité du modèle et de l'approche à être implémentés dans la réalité opérationnelle et à être validés par la fonction actuarielle.

C. Limites liées au contexte transactionnel

Lorsque l'on aborde la modélisation prédictive de la résiliation des contrats d'assurance dans un contexte transactionnel, plusieurs dimensions supplémentaires entrent en jeu. Ces aspects ajoutent des couches de complexité et soulèvent des questions spécifiques quant aux limites et aux défis à surmonter.

- **Contraintes temporelles dans les analyses transactionnelles :** le temps alloué pour l'analyse (*due diligences*) avant l'émission de l'offre peut être limité. Les décisions d'achat ou de vente doivent souvent être prises rapidement, ce qui impose une contrainte de temps sur l'analyse des données et la modélisation. Cette pression temporelle peut affecter la profondeur et la rigueur de l'analyse, conduisant potentiellement à des conclusions moins robustes et à un risque d'erreur inhérent.
- **Qualité et granularité des données transmises :** la qualité des données disponibles pour l'analyse est un facteur déterminant. Dans un cadre transactionnel, les données transmises peuvent varier en termes de granularité et de complétude. Des données de faible qualité ou

insuffisamment détaillées peuvent limiter la précision des modèles prédictifs et affecter la fiabilité des prévisions de résiliation.

- **Risques liés aux traitements "boîte noire" du Machine Learning :** l'utilisation de modèles de machine learning complexes, souvent perçus comme des "boîtes noires", peut introduire un risque d'erreurs non détectées. Dans un contexte transactionnel, où les enjeux sont élevés, il est crucial de comprendre comment les modèles génèrent leurs prédictions. Des erreurs ou des biais non identifiés dans les modèles peuvent conduire à des évaluations erronées de la valeur du portefeuille, avec des conséquences significatives sur les décisions transactionnelles.
- **Compétences requises au sein des équipes :** bien que les méthodes de machine learning soit devenus relativement connues scientifiquement, la complexité des modèles de machine learning nécessite des compétences spécialisées en data science et en analyse de données. Dans un contexte transactionnel, les équipes peuvent ne pas disposer de l'expertise requise pour développer, interpréter et valider de manière adéquate les modèles prédictifs. Cela soulève la question de la nécessité de faire appel à des consultants spécialisés pour mener à bien ces analyses ou de la formation des équipes internes ou le recours à des datascientists si l'entité acquérante a un nombre d'opportunités transactionnelles important. Les méthodes de machine learning devenant de plus en plus fiables et acceptées, il est raisonnable de s'attendre à ce que les professions évoluent pour intégrer les connaissances de base et fondamentales de ce domaine sur les années à venir.

Dans l'approche proposée, nous avons vu que le recours à la modélisation du risque de résiliation par machine learning s'avère pertinent dans un contexte transactionnel, par sa facilité de mise en œuvre à partir de la base de données client et contrats, et par sa faculté de prise en compte des comportements conjoncturels spécifiques à la clientèle du portefeuille étudié.

Cependant, en tenant compte des facteurs exposés ci-dessus (contraintes temporelles, qualité des données, risques liés aux modèles "boîte noire", compétences des équipes, etc.), il devient évident que la modélisation prédictive dans un contexte transactionnel présente des défis. Ces limites doivent être soigneusement prises en compte pour s'assurer que les analyses fournissent des résultats fiables et utiles pour guider les décisions.

II. Analyse des Données

1. Présentation du Dataset

Le dataset que nous avons utilisé (eusavingULnoPSperYr) est issu du package ‘CASdatasets’, une collection de données d'assurance compilée et maintenue par Arthur Charpentier et Christophe Dutang. Cette collection, initialement conçue pour le livre 'Computational Actuarial Science with R' édité par Arthur Charpentier, comprend désormais une large variété de datasets actuariels. Le package ‘CASdatasets’ est une ressource précieuse pour les chercheurs et les professionnels de l'actuariat, offrant un accès à des données réelles et pertinentes pour diverses applications.

Le dataset 'eusavingULnoPSperYr' est basé sur des données issues de produits d'épargne liés aux unités de compte (*unit-linked* en anglais), caractérisés par l'absence de participation aux bénéfices. Les données proviennent de plusieurs compagnies d'assurance vie anonymes de l'Union Européenne dont le pays européen est non spécifié. L'anonymat garantit la confidentialité et la protection des données, mais limite également la capacité à contextualiser les résultats dans le cadre spécifique de certaines compagnies ou marchés.

Nous considérons dans le cadre de notre étude qu’il s’agit ici d’un portefeuille d’assurance vie épargne en UC d’une seule compagnie d’assurance en prenant pour hypothèse de travail qu’il s’agit d’une population représentative de cet assureur précisément.

A. Des produits d'Épargne en Unités de Compte (Unit-Linked) :

Ces produits sont des formes d'assurance vie épargne où les bénéfices dépendent des performances des investissements sous-jacents. Ce produit d’investissement comporte un risque de perte en capital pour les assurés. Les souscripteurs investissent dans des unités de fonds, comme des fonds d'actions ou d'obligations, et la valeur de leur police fluctue en fonction des performances de ces investissements. Contrairement aux produits d'assurance vie traditionnels, il n'y a pas de participation aux bénéfices, ce qui signifie que les souscripteurs ne bénéficient pas directement des bénéfices réalisés par la compagnie d'assurance. Par ailleurs, ces produits comportent en général de meilleurs rendements que pour les produits ayant une participation aux bénéfices étant donné que le risque est ici porté par les assurés permettant une meilleure exposition au risque.

B. Période d'Observation (1999-2007)

Le dataset 'eusavingULnoPSperYr' couvre quasiment une décennie, de 1999 à 2007, une période qui permet d'observer les tendances et les comportements des souscripteurs sur un cycle économique complet, incluant des périodes de croissance et de récession.

En effet, voici quelques faits marquants qu'il convient d'avoir en tête pour notre étude au cours de cette période :

- **Fin des Années 1990 - Boom Économique :**
    - La fin des années 1990 a été caractérisée par une croissance économique robuste dans de nombreux pays européens. Cette période a été marquée par une forte expansion des marchés financiers, stimulée en partie par le boom des technologies de l'information et de la communication.
    - L'introduction de l'euro en 1999 a également joué un rôle clé, facilitant les échanges commerciaux et financiers entre les pays de la zone euro.
- **Début des Années 2000 - Ralentissement Économique :**
    - Au début des années 2000, l'économie européenne a connu un ralentissement, en partie dû à l'éclatement de la bulle Internet. Les marchés boursiers ont subi des corrections significatives, impactant les investissements et la confiance des consommateurs et des investisseurs.
    - Cette période a également été marquée par des incertitudes politiques et économiques, notamment après les attentats du 11 septembre 2001, qui ont entraîné des répercussions mondiales.
- **Milieu des Années 2000 - Reprise Économique :**
    - Vers le milieu des années 2000, l'économie européenne a commencé à se redresser. Cette reprise a été soutenue par une croissance mondiale solide, une augmentation des échanges commerciaux et une politique monétaire accommodante.
    - Cette période a également vu une augmentation significative des investissements immobiliers et une hausse des prix de l'immobilier dans plusieurs pays européens.
- **Crise Financière de 2007-2008 :**
    - La fin de la période couverte par le dataset coïncide avec le début de la crise financière mondiale de 2007-2008. Cette crise, déclenchée par l'effondrement du marché immobilier américain et la crise des subprimes, a entraîné des répercussions profondes sur les économies européennes.
    - Les marchés financiers ont connu une forte volatilité, et de nombreux pays ont été confrontés à des récessions, à une augmentation du chômage et à des défis budgétaires.

Ces différentes phases économiques ont eu un impact significatif sur les comportements des souscripteurs d'assurance vie en UC. En effet, les besoins de liquidité ou les craintes des assurés face aux fluctuations de économiques peuvent entraîner des pics de résiliations ou des pics de souscription. Les périodes de croissance pourraient encourager une plus grande prise de risque et d'investissement dans des produits liés aux unités, tandis que les périodes de récession pourraient entraîner une augmentation des résiliations dues à des contraintes financières ou à une aversion accrue au risque.

La période inclut des entrées (nouvelles souscriptions) et des sorties (résiliations) de polices, offrant une vision dynamique du marché de l'assurance vie en UC.

A noter que les données proviennent de compagnies d'assurance vie anonymes de l'Union Européenne. L'anonymat garantit la confidentialité et la protection des données, mais limite également la capacité à contextualiser les résultats dans le cadre spécifique de certaines compagnies ou marchés.

En analysant les données du dataset 'eusavingULnoPSperYr', nous pouvons donc espérer obtenir des éléments d'appréciations précieux sur la manière dont ces cycles économiques ont influencé les décisions des souscripteurs d'assurance vie épargne en UC.

C. Informations et caractéristiques disponibles dans le dataset

Le dataset 'eusavingULnoPSperYr' se distingue par sa richesse en informations, avec 30 colonnes couvrant divers aspects des contrats d'assurance et des indicateurs économiques :

- **Identifiants et Dates :**
  - **policy.ID** : Chaque police est unique et identifiable par son numéro, permettant un suivi précis sur toute la durée de l'étude.
  - **issue.date, termination.date :** Ces dates clés indiquent respectivement le début et la fin de chaque contrat, fournissant un cadre temporel pour chaque police.
- **Détails des Polices :**
  - **lapse.reason :** Cette variable cruciale renseigne sur les motifs de résiliation, offrant un aperçu direct des facteurs de désengagement des souscripteurs.
  - **premium.frequency, gender, underwriting.age :** Ces informations démographiques et contractuelles sont essentielles pour comprendre le profil et les préférences des souscripteurs.
  - **face.amount, risk.premium :** Elles donnent une idée de l'engagement financier des souscripteurs et de la structure des contrats.
- **Indicateurs Économiques et Financiers :** Des variables qui offrent un aperçu de l'environnement économique et financier, incluant l'évolution des indices de prix à la consommation, des indices boursiers, et des taux d'intérêt.

- **CPI.relvar** : Variation relative de l'Indice des Prix à la Consommation (CPI) sur différentes périodes.
- **EUidx.relvar :** Variation relative d'un indice boursier européen sur différentes périodes.
- **Rate1Y.relvar/ Rate2Y.relvar1/ Rate10Y.relvar1 :** Variation relative du taux d'intérêt à un an, deux ans, dix ans, sur différentes périodes.

- **D'autres indicateurs économiques et sociétaux :** Ces indicateurs reflètent les variations du marché du travail et du commerce de détail, fournissant un contexte économique plus large.
  - **unemploy.relvar** : Variation du taux de chômage européen.
  - **industry.relvar** : Variation d'un indice industriel européen.
  - **RTV.relvar :** Variation de l'indice du volume des ventes au détail européen.

D. Ajustement de la base de données :

Dans le cadre de notre étude visant à comprendre les motifs de résiliation des contrats d'assurance vie en UC et à développer un modèle prédictif de ces résiliations, nous avons apporté des modifications significatives au dataset issu de la collection CASdatasets.

Ces ajustements ont été conçus pour affiner notre analyse, en se concentrant sur les variables les plus pertinentes et en structurant les données de manière à faciliter l'interprétation et la modélisation.

- **Sélection et épuration des Variables :**

La première colonne, qui ne contenait que le numéro de ligne, ainsi que la colonne **premium.frequency**, ont été supprimées. Leur absence d'influence directe sur l'objectif de notre étude les rendait superflues pour la modélisation.

La variable **lapse.reason** a été transformée en variable binaire, où '1' représente une résiliation ('**Surrender'**) et '0' indique que la police est restée en vigueur (**'In Force'**). Cette codification simplifie les analyses multivariées et la construction de modèles prédictifs.

De même, le genre du souscripteur (**gender**) a été encodé en valeurs binaires, avec '1' pour les hommes et '0' pour les femmes, afin de standardiser les données pour les analyses ultérieures.

- **Ajout et transformation de Variables :**

Un nouveau champ, **Tenure**, a été introduit pour refléter l'ancienneté de la relation avec le client. Cette variable augmente annuellement tant que le client reste en portefeuille, offrant ainsi une dimension temporelle précieuse à notre analyse.

La colonne **termination.date** a été retirée pour éviter de donner des indices trop explicites qui pourraient biaiser la prédiction de résiliation (information évidente).

Les informations contenues dans **issue.date** ont été décomposées en deux colonnes distinctes, **issue.year** et **issue.month**, permettant une analyse plus granulaire du temps en relation avec la dynamique de résiliation.

- **Déduction de l'âge de la police :**

En utilisant l'âge de souscription (**underwriting.age**) et l'année d'entrée en relation (**issue.year**), nous avons calculé l'âge de la police à chaque observation, ce qui fournit un indicateur dynamique de la maturité de la relation entre le client et l'assureur.

- **Calcul de la date d'observation :**

À partir de l'année d'entrée en relation (**issue.year**), nous avons déterminé la date d'observation des paramètres pour chaque ligne, ce qui permet de contextualiser chaque donnée dans son cadre temporel spécifique.

Ces modifications méthodiques transforment le dataset initial en un outil plus affiné pour notre analyse, en éliminant les redondances et en clarifiant les variables clés. L'introduction de Tenure et la transformation des variables catégorielles en formats binaires sont des étapes cruciales pour préparer les données à des analyses multivariées et à la modélisation prédictive. En retirant les variables potentiellement prédictives mais trop évidentes dans le cadre de prédiction, comme **termination.date**, nous nous assurons que le modèle se concentrera sur les facteurs sous-jacents influençant la décision de résiliation, plutôt que sur des indicateurs post facto.

La préparation minutieuse du dataset est une étape fondamentale de notre recherche, visant à isoler les facteurs influençant les résiliations et à construire un modèle prédictif robuste et éclairé.

Ces ajustements méthodologiques permettent d'aborder l'analyse avec une base de données optimisée, prête à révéler les dynamiques complexes à l'œuvre dans les décisions de résiliation des contrats d'assurance vie en UC.

2. Analyse Exploratoire des Données (EDA)

A. Analyse du portefeuille

Il est primordial de comprendre la composition du portefeuille avant de se lancer dans les analyses exploratoires de données plus avancées pour comprendre le comportement de la clientèle qui compose ce portefeuille en fonction des facteurs économiques, financiers et sociétaux. Analyser la composition du portefeuille implique d'examiner des aspects démographiques tels que l'âge, le genre, la fréquence de paiement des primes, et d'autres variables qui peuvent influencer la fidélité et la rentabilité des clients. Le dataset renseigne une partie de ces informations.

Dans le cadre d'une transaction, l'acquéreur potentiel évalue méticuleusement la base de clientèle de l'entité cédante. Cette évaluation n'est pas superficielle ; elle cherche à appréhender la valeur intrinsèque et le potentiel de rentabilité futur du portefeuille. La base de clientèle est souvent perçue comme un actif précieux, car elle incarne non seulement les flux de revenus actuels mais aussi les perspectives de croissance et de stabilité financière à long terme.

Ainsi pour un acquéreur, la compréhension approfondie du portefeuille est synonyme de prise de décision éclairée. Elle lui permet d'évaluer le risque associé à l'investissement et de déterminer la valeur réelle qu'il est prêt à attribuer à la base de clientèle. Cette évaluation influence directement le prix d'achat et les conditions de la transaction, soulignant ainsi l'importance stratégique de l'analyse préalable du portefeuille.

Nous commençons par analyser la base de clientèle au niveau de l'âge de souscription des assurés, ce qui devrait refléter des informations précieuses au regard des moments de vie différents des assurés. A noter qu'il s'agit d'une variable importante pour la modélisation et qu'une étude par tranche d'âge plus fine serait pertinente.

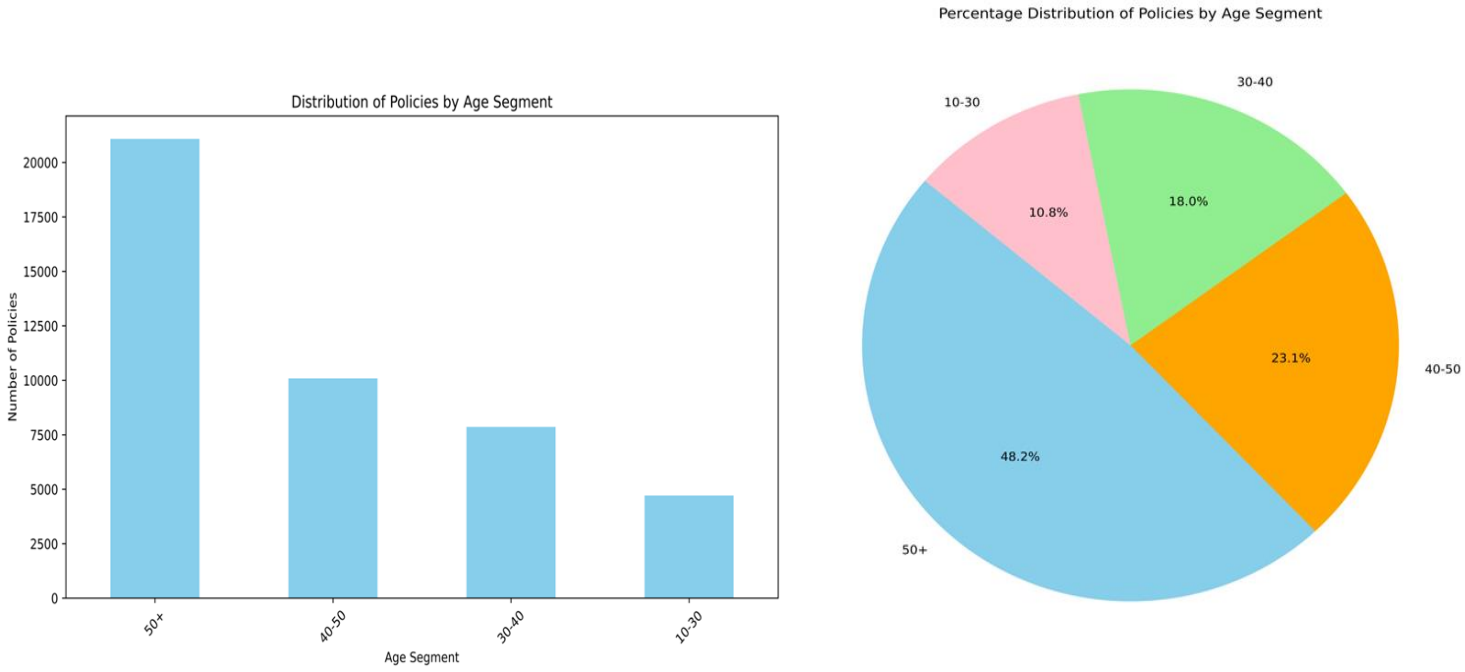


Figure : Nombre de police et répartition (%) par tranche d'âge

- **Prédominance des segments d'âge plus élevés** : Le segment d'âge "50+" domine nettement la distribution, indiquant que les individus dans cette tranche d'âge sont les plus enclins à souscrire des polices d'assurance. Cela peut refléter une prise de conscience accrue des besoins de sécurité financière et de protection à mesure que les individus vieillissent.

- **Souscription progressive avec l'âge :** On observe une augmentation progressive du nombre de polices souscrites à mesure que l'âge augmente, de "10-30" à "50+". Cela suggère que la tendance naturelle à investir dans des polices d'assurance croît avec l'âge, probablement en raison de l'évolution des priorités financières et de la prise de conscience des risques associés à l'âge.
- **Moindre engagement des jeunes :** Le segment "10-30" présente le nombre le plus faible de polices souscrites, ce qui pourrait indiquer un manque de sensibilisation ou d'intérêt pour l'assurance vie parmi les jeunes adultes. Cela peut également refléter des contraintes budgétaires ou une perception de moindre besoin de protection à cet âge.

Cette distribution souligne l'importance pour les assureurs de cibler les segments d'âge plus âgés avec des produits adaptés à leurs besoins spécifiques de sécurité financière et de protection. Toutefois, elle met également en évidence une opportunité pour les assureurs de sensibiliser davantage les plus jeunes à l'importance de la souscription précoce à des polices d'assurance, potentiellement en adaptant les produits et les stratégies de communication pour mieux répondre à leurs besoins et préférences.

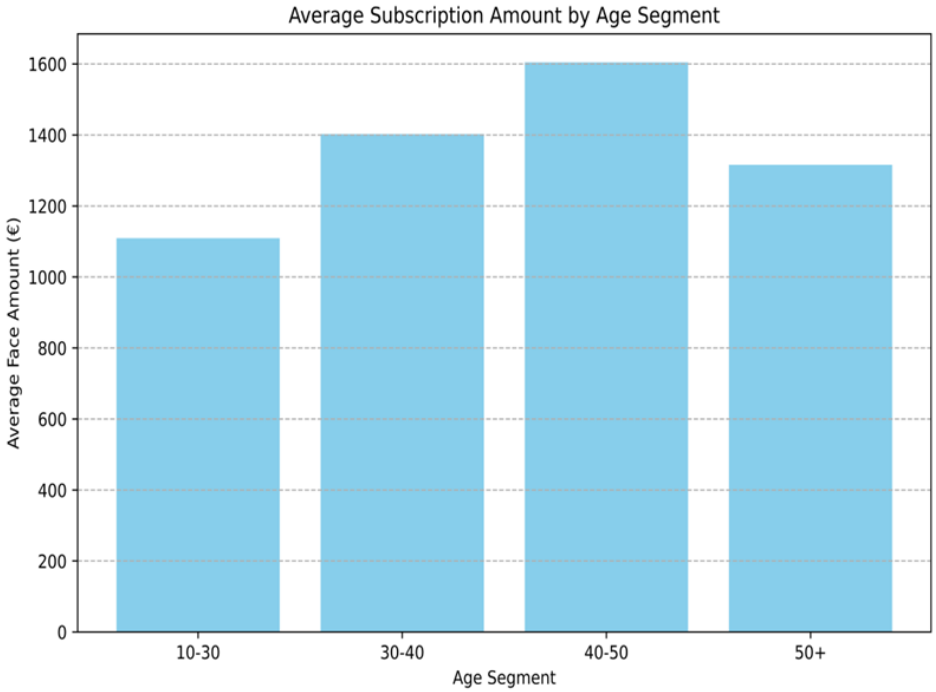


Figure : Montant moyen des souscriptions par tranche d'âge

Le manuel du dataset ne fournit pas d'information précise quant à ce que représente le face.amount. Par simplification pour notre étude, nous supposerons qu'il s'agit du niveau d'encours de la police au moment de la souscription.

Les montants moyens de souscription semblent relativement faibles, ce qui ne semble pas représentatif de la population française.

- **Segment 10-30 ans :** Le montant moyen de souscription est de 1 109,29 €. Ce montant plus faible pourrait refléter une capacité financière limitée ou une perception de moindre besoin de couverture d'assurance dans cette tranche d'âge.
- **Segment 30-40 ans :** Le montant moyen augmente à 1 402,59 €, ce qui peut indiquer une amélioration de la situation financière ou une prise de conscience accrue des besoins d'assurance à mesure que les individus avancent dans leur carrière et commencent à fonder des familles.
- **Segment 40-50 ans :** Le montant moyen atteint son pic à 1 604,57 €, suggérant une capacité financière supérieure malgré un besoin plus élevé de responsabilité financière, comme l'éducation des enfants ou le remboursement d'emprunts immobiliers. Cette croissance de la capacité financière par rapport aux plus jeunes âges semble cohérente de par la stabilité financière. En effet, les personnes âgées de 40 ans et plus sont souvent dans une phase plus stable de leur vie professionnelle et financière. Elles sont plus susceptibles d'avoir des revenus réguliers et élevés, ce qui les rend capables de maintenir leurs polices d'assurance sur le long terme, réduisant ainsi le risque de résiliation pour l'assureur.
- **Segment 50+ ans :** le montant moyen diminue légèrement à 1 315,99 €. Cela pourrait refléter une transition vers la retraite avec une réduction des revenus disponibles ou une réévaluation des besoins d'assurance à mesure que les obligations financières majeures, comme l'éducation des enfants, sont complétées. Cela peut aussi refléter qu'il s'agisse d'une clientèle « mass market » dont les personnes au-delà de 50+ ans ne disposent pas forcément de patrimoine conséquent qui nécessiterait un besoin de placements important.

La prédominance des segments d'âge plus élevés (40-50 et 50+) présente plusieurs avantages clés pour un assureur. En effet, les individus dans ces tranches d'âge ont tendance à avoir des montants d'encours supérieurs, comme le montre l'analyse des montants moyens de souscription. Cela signifie que les primes versées sont plus élevées, ce qui peut générer des revenus plus importants pour l'assureur à travers les frais de gestion des contrats et des unités de compte.

Aussi, les personnes âgées de 40 ans et plus sont souvent dans une phase plus stable de leur vie professionnelle et financière. Elles sont plus susceptibles d'avoir des revenus réguliers et élevés, ce qui les rend capables de maintenir leurs polices d'assurance sur le long terme.

Avec l'âge, la prise de conscience des besoins de planification financière pour la retraite ou la transmission de patrimoine s'accroît. Cela rend les produits d'assurance épargne en UC particulièrement attrayants pour ces segments d'âge, car ils offrent à la fois une opportunité de croissance du capital et des avantages fiscaux.

Ces clients sont par ailleurs souvent à la recherche de diversification de leurs investissements pour optimiser leur rendement et minimiser les risques. Les assurances épargne en UC permettent cette diversification à travers une large gamme de supports d'investissement, répondant ainsi à leurs besoins.

Ils représentent aussi un potentiel de fidélisation élevé pour l'assureur. En proposant des services et des conseils personnalisés, l'assureur peut renforcer la relation avec ces clients, les incitant à maintenir leurs investissements et à explorer d'autres produits d'assurance.

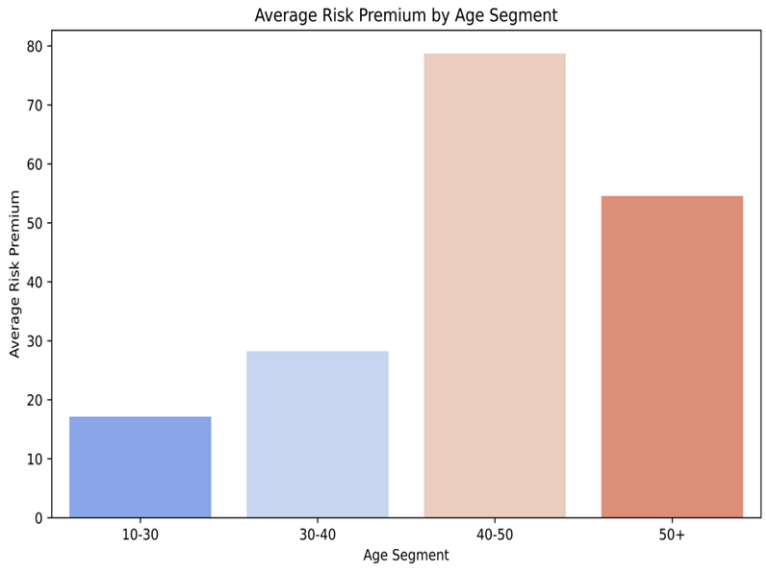


Figure : Prime de risque (profil de risque) par tranche d'âge

Le manuel du dataset ne fournit pas d'information précise quant au calcul des primes de risques. Nous supposerons qu'il s'agit d'une cotation du niveau d'exposition de risque de la police.

Les personnes plus âgées présentent un profil de risque plus élevé par rapport aux jeunes souscripteurs. Cette observation peut s'expliquer par la sensibilité aux événements de marché. Les souscripteurs plus âgés sont souvent plus exposés à des produits financiers indexés sur les marchés, tels que les fonds de pension ou les plans d'épargne en actions, dans le cadre de leur planification de retraite. Cette exposition plus importante aux fluctuations du marché peut rendre leurs polices plus sensibles aux événements de marché, ce qui se reflète dans des primes de risque plus élevées.

Nous étudions ensuite les différentes phases de collecte et de résiliation sur la période étudiée. Il est important de mentionner que le dataset ne contient les entrées en relation et les résiliations qu'entre 1999 et 2007. De plus, le niveau d'encours mentionné dans la base n'évolue pas pour la même police ; nous pouvons supposer qu'il s'agit du niveau d'encours observé à la fin de la période d'observation ou au moment de la résiliation si la police a résilié son contrat. Par souci de simplification de notre analyse qui porte sur l'estimation du risque de résiliation, nous considèrerons que le niveau d'encours est celui à la souscription et qu'il n'y a pas de versement supplémentaire au cours du contrat.

Voici quelques remarques :

- **Accumulation progressive :** Les encours totaux augmentent au fil des années du fait de l'accumulation progressive des montants dus par les polices émises au cours des différentes années. Cela reflète à la fois l'ajout de nouvelles polices sur la période étudiée et le fait que les entrées sont plus importantes que les sorties, ce qui est tout à fait attendu étant donné qu'il s'agit d'un portefeuille en premières phases de constitution pour notre étude. Les encours se stabiliseront lorsque le niveau de sorties sera plus ou moins équivalent aux entrées, définissant ainsi un niveau de portefeuille mature. Il conviendrait évidemment d'avoir l'ensemble du portefeuille de cet assureur avec une meilleure profondeur de l'historique pour analyser plus précisément l'évolution totale du portefeuille concerné, mais nous sommes contraints dans le cadre de cette étude de considérer ce portefeuille ci, en cours de constitution.
- **Variabilité des contributions :** Il y a une variabilité dans la collecte par années d'émission. Certaines années semblent avoir eu un impact plus significatif sur l'augmentation des encours totaux, ce qui pourrait être dû à une augmentation du nombre de polices émises ou à des conditions de marché favorables. Cela peut aussi indiquer une stratégie d'expansion ou une réponse à une demande croissante du marché.

Un autre critère qu'il nous parait pertinent d'observer est le lien entre le taux de résiliation et l'ancienneté de la police :

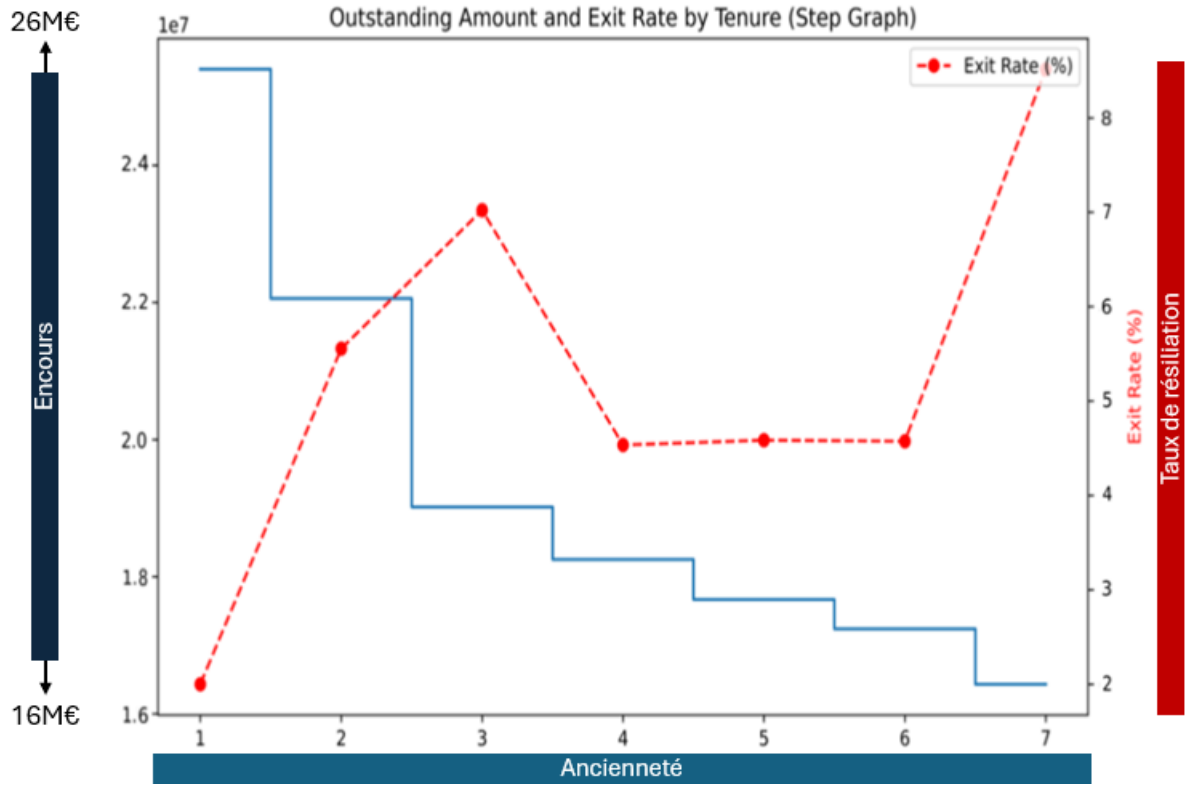


Figure : Evolution du niveau de l'encours selon l'ancienneté de la police (« Tenure »)

Nous observons un déséquilibre des classes, avec une sur-représentation des polices n'ayant pas résilié par rapport à celles ayant résilié (phénomène d'imbalance). Les résiliations sont de l'ordre de 5% à 8% annuel selon l'ancienneté). Dans ce contexte, les modèles prédictifs risquent de favoriser la classe majoritaire afin de minimiser leur erreur globale, ce qui peut entraîner une sous-estimation systématique des résiliations. Un modèle entraîné sur des données déséquilibrées aura ainsi tendance à ignorer les schémas comportementaux spécifiques aux assurés les plus susceptibles de résilier, compromettant la fiabilité des prévisions. Pour atténuer cet effet, plusieurs stratégies peuvent être mises en place, telles que le rééchantillonnage des données par oversampling de la classe minoritaire ou l'ajustement des poids associés aux classes lors de l'apprentissage du modèle. Ces techniques permettent de rééquilibrer les prédictions et d'améliorer la capacité du modèle à détecter les résiliations, tout en conservant une bonne généralisation aux données futures.

Par ailleurs, il est important de mentionner que la profondeur des données (8 ans) est dans les faits insuffisante pour traiter la prédiction de résiliation d'un contrat d'assurance vie dont la durée moyenne en France est de l'ordre de 13 ans selon les données de France Assureurs. A noter que les mesures fiscales incitent les assurés à conserver son contrat pendant au moins 8 ans afin de bénéficier d'une fiscalité plus avantageuse.

Ce graphique en escalier offre une perspective visuelle intuitive sur la dynamique des encours et des taux de résiliation au fil du temps, mettant en évidence les tendances claires.

Nous constatons en effet un niveau important de résiliations durant les premières années, particulièrement jusqu'à la troisième année. Cette tendance peut être interprétée de deux manières principales. Premièrement, elle pourrait indiquer que certains investisseurs optent pour des placements à court terme, envisageant dès le départ une stratégie d'investissement de durée limitée, qui pourrait éventuellement s'expliquer par un avantage fiscal particulièrement sensible trois ans ou 7 ans après leurs investissements initiaux. Deuxièmement, cette observation pourrait refléter l'impact d'événements spécifiques sur les marchés financiers, qui ont conduit à une augmentation des taux de résiliation parmi les souscripteurs.

Cette tendance suggère également une corrélation avec les cycles économiques, marquant une alternance entre les phases de reprise et de récession. Les pics de résiliation observés pourraient ainsi coïncider avec des périodes économiques défavorables, où les investisseurs, face à l'incertitude ou à la baisse des rendements, choisissent de se retirer. Cette dynamique souligne l'importance de la prise en compte du contexte économique global et de ses fluctuations lors de l'analyse des comportements d'investissement dans le secteur des assurances vie en UC.

Nous décidons d'illustrer comment les taux de sortie varient en fonction de l'ancienneté des encours et de l'âge de souscription, ce qui peut offrir une perspective sur la manière dont ces facteurs peuvent influencer les comportements de sortie des encours.

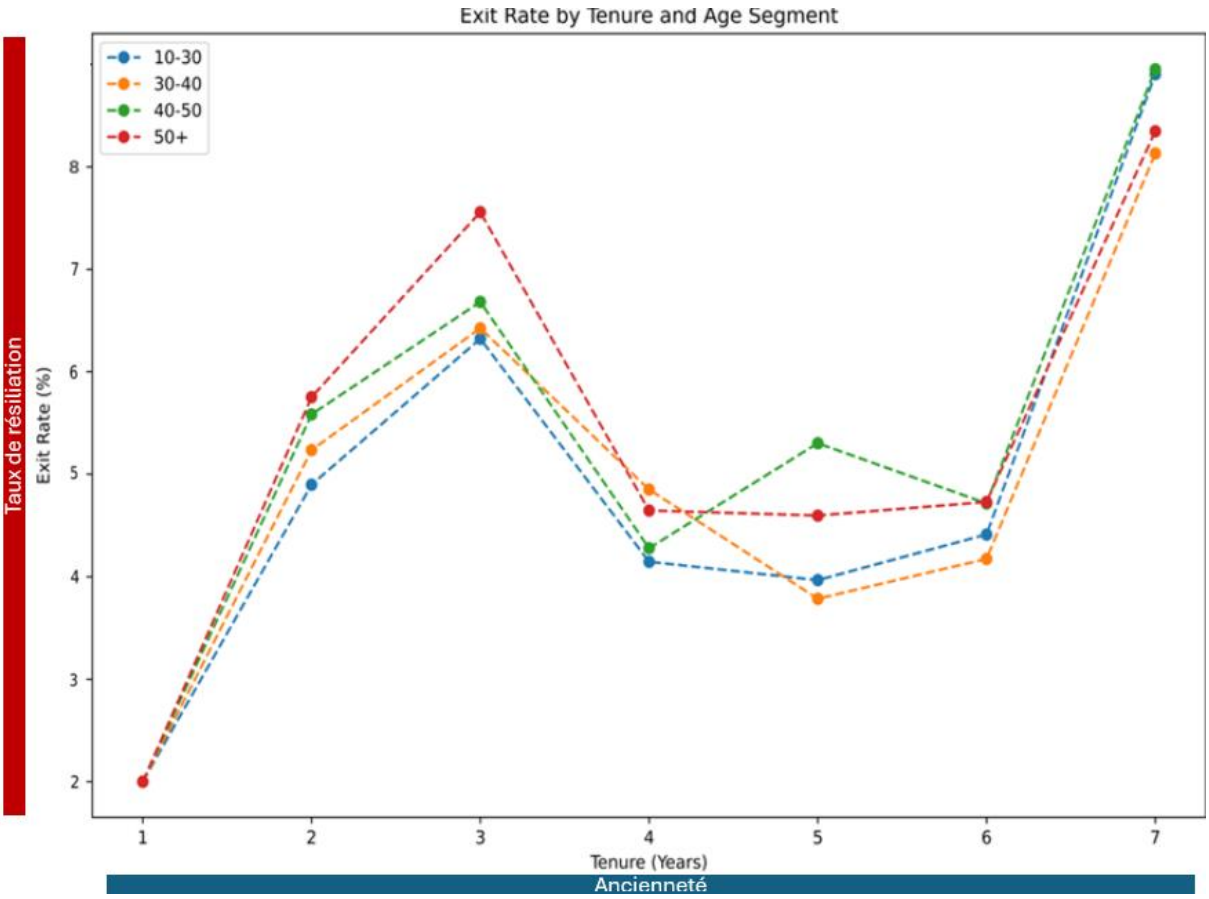


Figure : Evolution du taux de sortie des encours selon l'ancienneté de la police par tranche d'âge

Nous pouvons observer une légère variation des taux de sortie en fonction de la segmentation d'âge. Cela indique que l'âge au moment de la souscription peut influencer la probabilité de sortie des encours au fil du temps.

Nous constatons une corrélation entre l'ancienneté et le taux de rachat qui est frappante, avec un pic à 3 ans et à 7 ans.

Certaines tranches d'âge peuvent présenter des taux de sortie plus élevés ou plus volatils que d'autres. Cela peut refléter des différences dans les comportements financiers ou les besoins de liquidité entre les groupes d'âge, surement corrélées aux différents moments de vie.

L'observation selon laquelle les personnes âgées de 10 à 30 ans ont tendance à moins résilier leurs encours par rapport aux populations plus âgées pourrait s'expliquer par plusieurs facteurs socio-économiques et comportementaux :

- **Enrichissement en cours :** Les individus dans la tranche d'âge de 10 à 30 ans sont souvent en début de carrière professionnelle et peuvent ne pas avoir accumulé suffisamment de richesse ou d'actifs pour se sentir à l'aise de résilier leurs encours. Ils peuvent être plus enclins à conserver leurs investissements ou leurs polices d'assurance pour la sécurité financière à long terme, d'autant plus que leur niveau d'investissement est plus faible que pour les populations plus âgées.
- **Placements moins impactés par des besoins de liquidité :** Les personnes plus jeunes peuvent avoir moins de dépenses importantes immédiates auxquelles devoir faire face (telles que les frais de scolarité des enfants ou les dépenses de santé liées à l'âge) par rapport aux tranches d'âge plus âgées, réduisant ainsi leur besoin de liquidités et la probabilité de résilier leurs encours.
- **Influence des moments de vie :** Les événements de vie majeurs (comme l'achat d'une maison, le mariage, ou la naissance d'enfants) qui peuvent inciter à la résiliation d'encours pour financer ces événements sont plus fréquents à mesure que les gens vieillissent.
- **Connaissance et expérience financières :** Les individus plus âgés peuvent avoir une meilleure compréhension des marchés financiers et des produits d'assurance, ainsi qu'une plus grande expérience dans la gestion de leurs finances, ce qui peut les amener à prendre des décisions plus actives concernant la résiliation de leurs encours pour une réallocation vers d'autres supports d'investissement par exemple.

Ces explications sont évidemment générales et peuvent varier en fonction des circonstances individuelles, des conditions de marché, et des caractéristiques spécifiques des produits d'encours.

B. Statistiques descriptives : analyse univariée

L'examen initial de la qualité des données contenues dans le dataset 'eusavingULnoPSperYr' révèle des résultats encourageants pour la suite de notre étude. Cette première étape d'analyse est cruciale pour assurer la fiabilité et la validité des conclusions tirées ultérieurement.

Une caractéristique notable du dataset est l'absence de valeurs manquantes. Cette complétude des données est essentielle, car les valeurs manquantes peuvent souvent introduire des biais dans l'analyse et compliquer l'interprétation des modèles prédictifs. Le fait que toutes les entrées soient complètes signifie que nous disposons d'un ensemble de données cohérent et intégral pour notre analyse.

L'absence de lignes dupliquées dans le dataset est également un point positif. Les doublons peuvent fausser les analyses statistiques et les modèles prédictifs en sur-représentant certaines informations. Leur absence garantit que chaque entrée dans notre dataset représente une observation unique, ce qui est crucial pour l'intégrité de notre étude.

La réalisation d'un résumé statistique des variables numériques du dataset est une étape importante. Ce résumé fournit une vue d'ensemble des caractéristiques clés de ces variables, telles que la moyenne, la médiane, les écarts-types, et les quartiles.

Les statistiques descriptives pour les variables clés montrent les informations suivantes :

- **lapse.reason :** Environ 41% des polices ont résilié (valeur 1 pour Surrender, 7632 polices), tandis que 59% sont encore en vigueur (valeur 0 pour In Force, 10 933 polices).
- **gender :** Environ 62% des souscripteurs sont des hommes (valeur 1), contre 38% de femmes (valeur 0).
- **underwriting.age :** L'âge moyen lors de la souscription est d'environ 49 ans, avec un âge minimum de 14 ans et un maximum de 99 ans. Ces résultats montrent que l'âge moyen lors de la souscription est assez similaire entre les hommes et les femmes, et ne varie pas significativement entre ceux qui ont résilié leurs polices et ceux qui ne l'ont pas fait.

Gender	Lapse Reason	Average Underwriting Age
Femme	Non résiliée	49.91
Femme	Résiliée	49.74
Homme	Non résilié	48.80
Homme	Résilié	48.78

- **face.amount :** Le montant moyen de la couverture est d'environ 1 375,86, avec un minimum de 1,41 et un maximum de 88 953,79€.
- **Tenure :** L'ancienneté moyenne des polices est d'environ 2,2 ans, avec une durée minimale de 1 an et une maximale de 8 ans (sachant que le dataset ne nous permet pas de visualiser au-delà de 8 ans).

Le résumé statistique est particulièrement utile pour identifier les éventuelles valeurs aberrantes. Ces valeurs aberrantes peuvent être le résultat d'erreurs de saisie, de mesures exceptionnelles ou de variations naturelles. Leur identification est essentielle, car ils peuvent avoir un impact significatif sur les résultats de l'analyse et sur la performance des modèles prédictifs. Après analyse, ces valeurs nous apparaissent comme justes. Des extrêmes existent mais ils ne représentent pas forcément des valeurs aberrantes.

Dans le cadre de l'analyse descriptive, nous observons les distributions des variables numériques pour comprendre la représentativité des données de clients qui composent le dataset étudié.

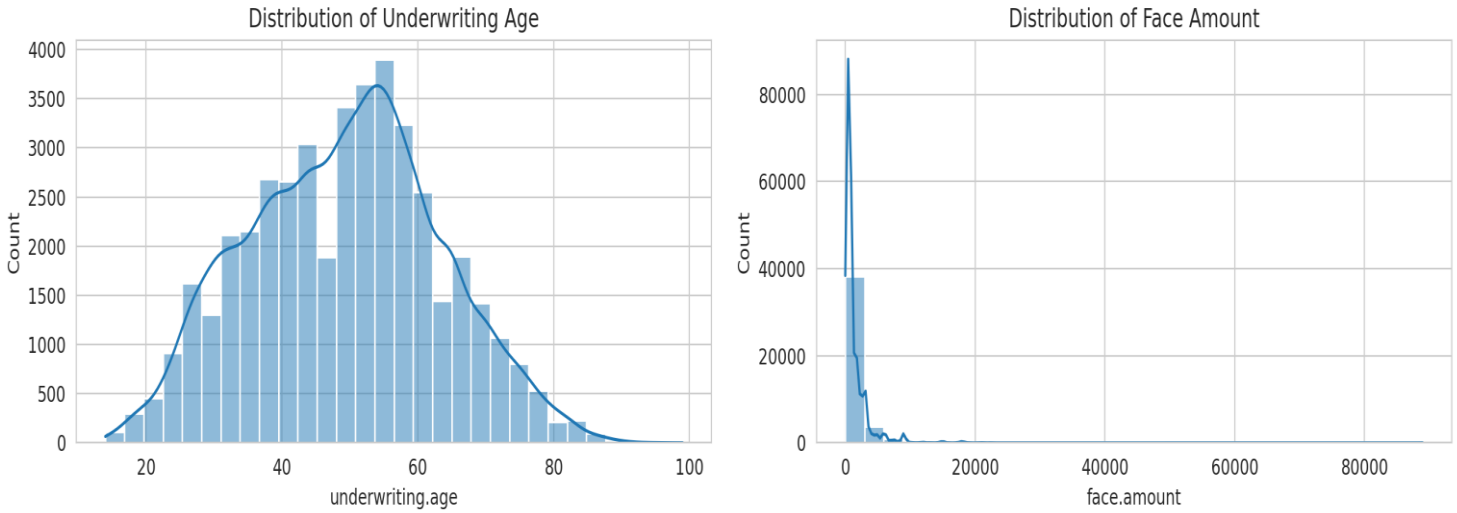


Figure : Divers distributions des données disponibles (Univariate analysis)

- **Age de souscription :** La distribution semble être centrée autour de la cinquentaine avec une queue plus longue vers les âges plus élevés.
- **Montant assuré (Face amount) :** La distribution du montant assuré montre une concentration élevée pour les valeurs plus faibles, indiquant que la plupart des polices ont des montants assurés relativement modestes, ce qui témoigne d'un segment de clientèle étudié « mass market ». Étant donné la large gamme de valeurs, une échelle logarithmique est utilisée pour l'axe des x, ce qui permet de mieux visualiser la distribution des montants plus faibles, qui sont les plus fréquents.

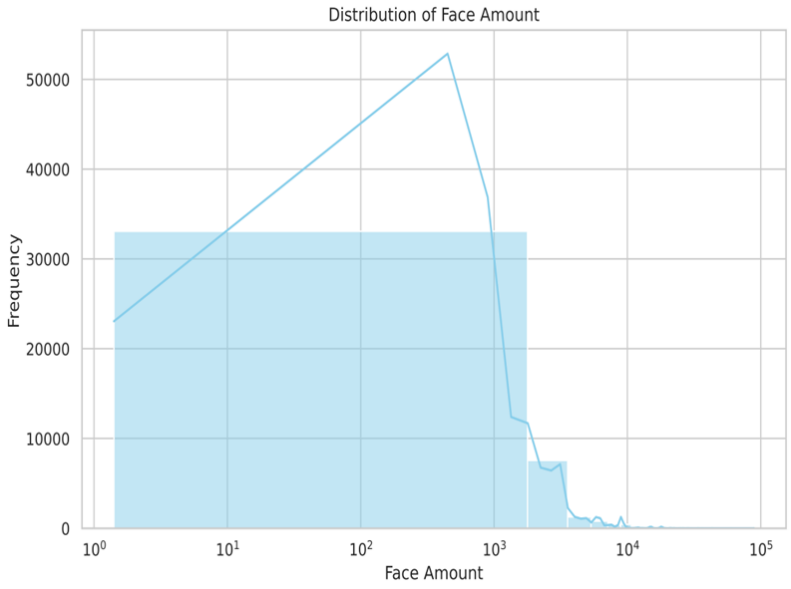


Figure : distribution des montant assurés (échelle logarithmique)

L'augmentation du nombre d'intervalles utilisées pour afficher la distribution des montants assurés dans le graphique logarithmique permettrait de mieux visualiser la répartition des montant. Par ailleurs, la

majorité des valeurs sont regroupées à gauche ce qui implique beaucoup de petits contrats et très peu de très gros contrats (distribution fortement asymétrique).

Les primes de risques faibles conjuguées aux primes d'épargne faibles et aux montants assurés faibles démontrent qu'il s'agit d'un portefeuille de clientèle plutôt classique se rapprochant très probablement d'un segment de clients « mass market ».

Figure : Distribution des contrats par ancienneté

- **Ancienneté de la police (Tenure) :** La distribution de la durée des polices montre une concentration pour les durées plus courtes, indiquant que de nombreuses polices sont résiliées ou arrivent à échéance relativement tôt. La majorité des polices ont une ancienneté de 1 à 3 ans.

Nous allons maintenant observer les distributions des indicateurs économiques et financiers ainsi que des autres indicateurs économiques et sociétaux.

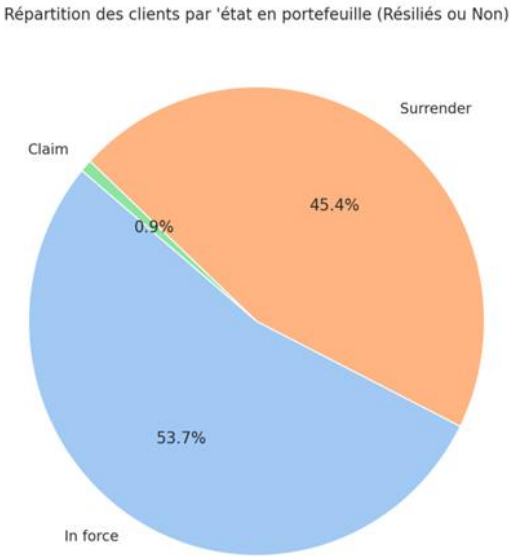
Voici quelques observations générales basées sur les tendances visuelles :

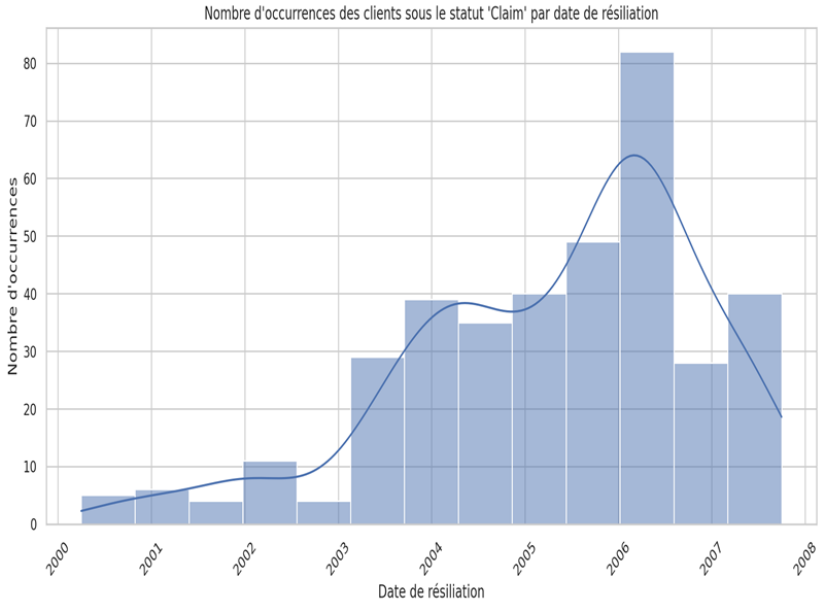
- **Variation du CPI (CPI.relvar) :** Des pics et des creux sont observables et pourraient indiquer des périodes de forte inflation ou de déflation, souvent déclenchées par des crises économiques, des chocs pétroliers, ou des politiques monétaires.
- **Indice économique de l'UE (EUidx.relvar)** : Des fluctuations sont visibles, suggérant des périodes de croissance économique et de récession au sein de l'Union Européenne.
- **Taux d'intérêt (Rate1Y.relvar, Rate2Y.relvar, Rate10Y.relvar) :** Les variations des taux d'intérêt sur 1 an, 2 ans et 10 ans montrent des changements dans la politique monétaire qui

pourraient refléter des réponses à des conditions économiques changeantes. Par exemple, une baisse des taux peut indiquer une tentative de stimuler l'économie pendant une récession.

- **Taux de chômage (unemploy.relvar) :** Des variations significatives pourraient indiquer des périodes de crise économique ou de reprise. Une augmentation soudaine du taux de chômage peut être liée à des récessions économiques ou à des crises industrielles, tandis qu'une diminution peut indiquer une reprise économique.
- **Indice de l'industrie (industry.relvar) :** Les changements dans cet indicateur peuvent refléter l'état de la production industrielle et de l'activité économique globale. Les fluctuations dans ce secteur peuvent être influencées par des changements dans la demande globale, des innovations technologiques, ou des politiques commerciales.
- **Volume des ventes au détail (RTV.relvar) :** Des variations ici pourraient indiquer des changements dans la confiance des consommateurs et dans les habitudes de dépense. En effet, le commerce de détail est souvent affecté par le pouvoir d'achat des consommateurs, qui peut fluctuer en fonction de la santé économique globale, des niveaux de confiance des consommateurs, et des politiques fiscales.

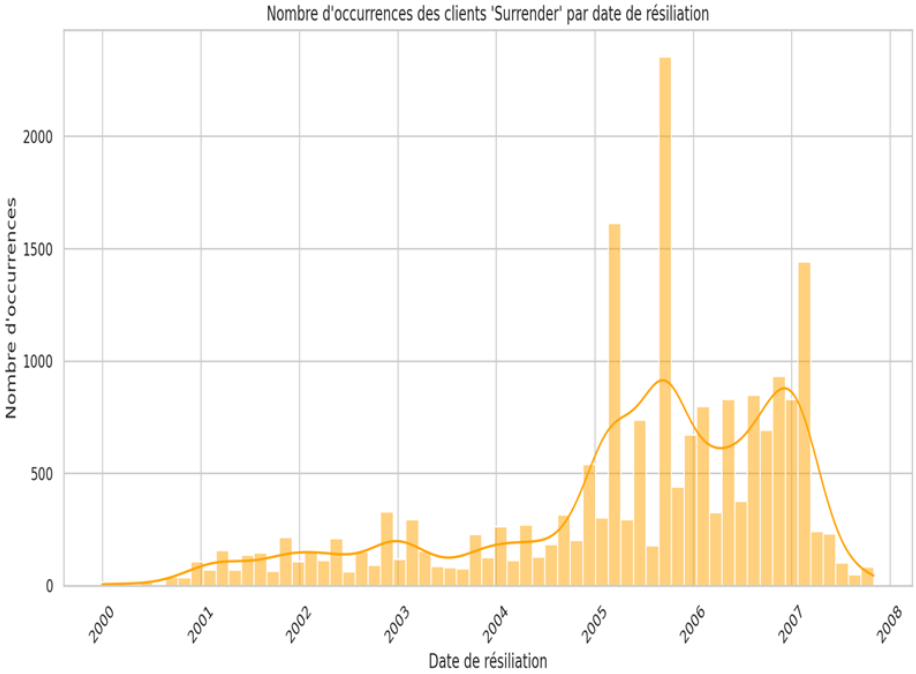
Voici la répartition des clients selon leur état en portefeuille, indiquant s'ils sont résiliés ou non :





Les Claims (rachats partiels) étant peu représentatifs, nous décidons de considérer les contrats concernés encore en portefeuille (in force) pour notre étude.

Voici le nombre d'occurrences des clients avec le statut "Surrender" (résiliés) par date de résiliation :



Nous pouvons observer des pics qui pourraient correspondre à des périodes spécifiques où un nombre plus élevé de clients a choisi de résilier leurs polices. Le pic de résiliation se situe entre 2005 et 2007 lors de période de reprise économique soutenue par une croissance mondiale globalisée. Cette période a également vu une augmentation significative des investissements immobiliers et une hausse des prix de l'immobilier dans plusieurs pays européens.

Importance du genre concernant la résiliation

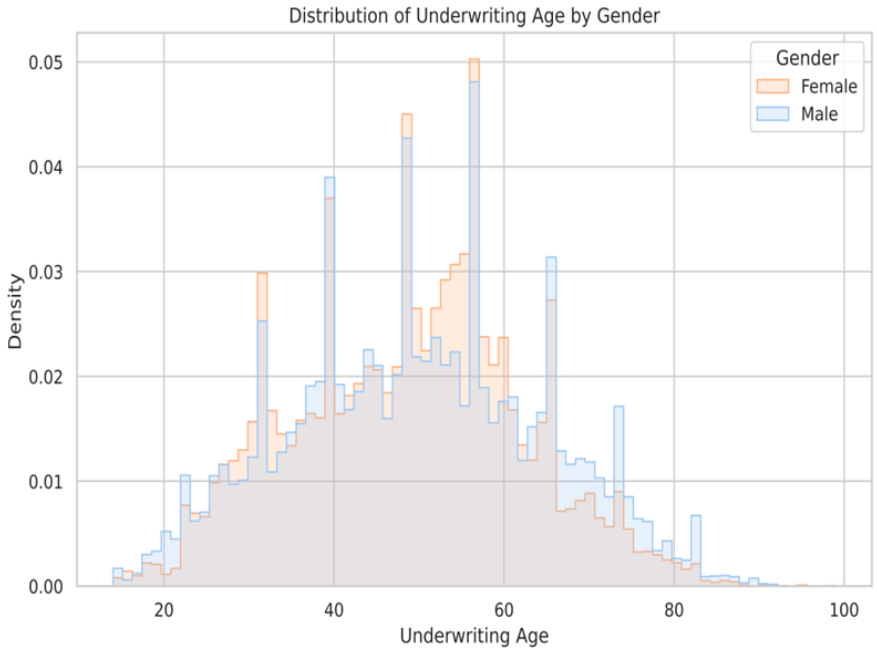


Figure : Distribution de l'âge lors de la souscription par genre

Cette visualisation montre la distribution de l'âge lors de la souscription pour les hommes et les femmes. Les deux distributions semblent assez similaires, avec une légère prédominance féminine dans la tranche d'âge moyenne et masculine dans les tranches d'âges supérieures à 60 ans.

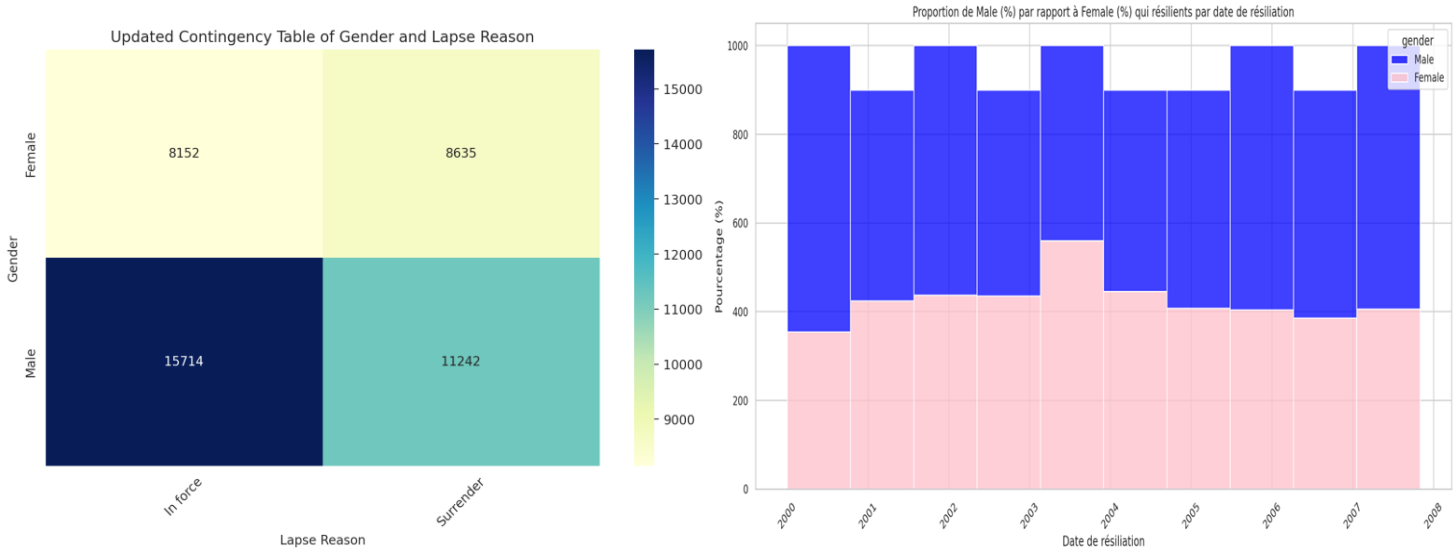
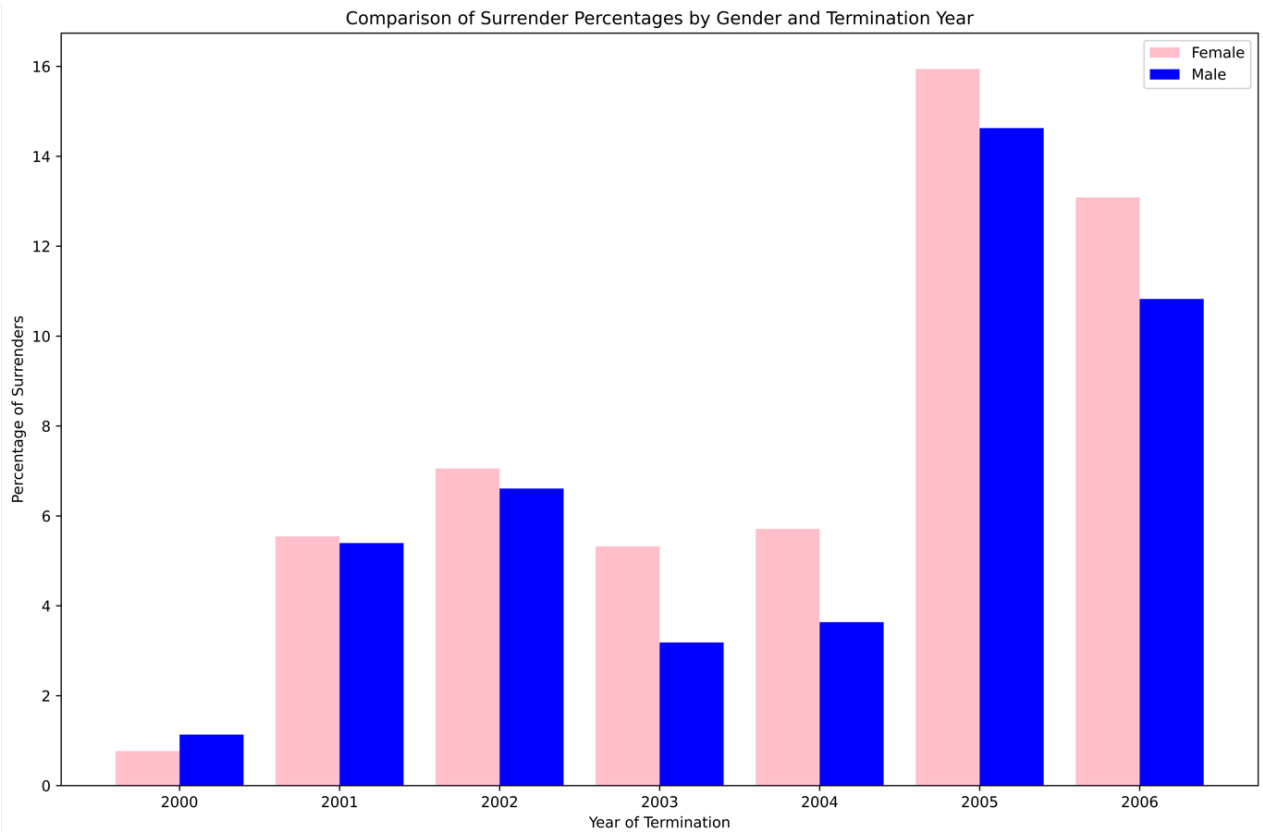


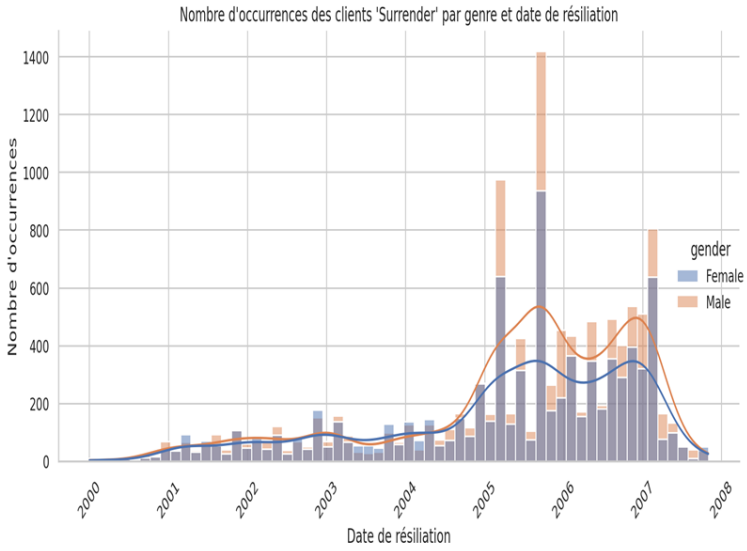
Figure : Comparaison des occurrences selon les genres et si résiliés ou non

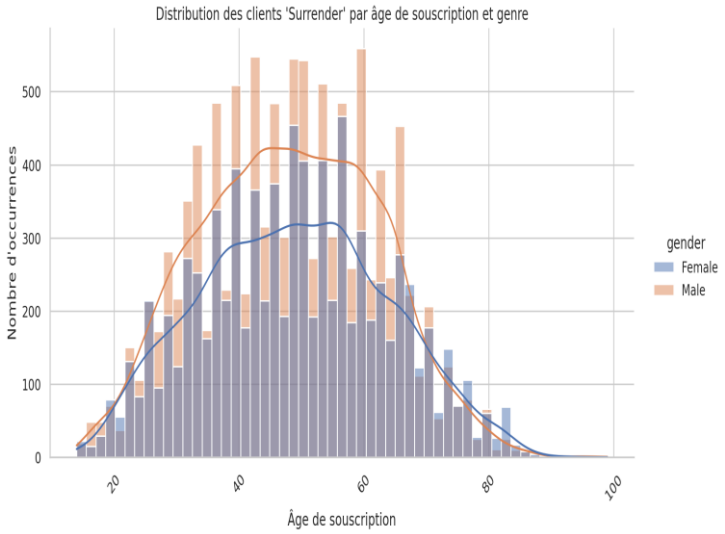
Sur le graphique Heatmap (à gauche), la couleur indique la fréquence des occurrences, permettant de visualiser l'association entre le genre et si la police est résiliée ou non. Nous constatons qu'il y a une sur-représentation des hommes dans le dataset.

A droite, l'histogramme illustre la proportion de clients de genre masculin (en bleu) et féminin (en rose) qui ont résilié leur police d'assurance, répartie par date de résiliation. Cela permet de visualiser la dynamique de résiliation entre les genres au fil du temps. Nous constatons que la proportion des femmes par rapport aux hommes dans les résiliations est quasiment du même ordre. Les hommes étant sur-représentés dans la base, nous pouvons en déduire que les femmes comportent un risque de résiliation de leur contrat d'assurance vie épargne UC plus important que les hommes.

Le graphique suivant qui présente le comparatif des pourcentages de résiliation par genre et par année de résiliation conforte cette hypothèse :







- **Tendances Générales :** On observe des fluctuations dans les taux de résiliation pour les deux genres au fil des années. Ces variations peuvent refléter des changements dans les conditions du marché, des politiques internes de l'entreprise d'assurance, ou des facteurs socio-économiques externes influençant les décisions des assurés.
- **Comparaison entre les Genres :** Il est intéressant de noter que, pour certaines années, les pourcentages de résiliation diffèrent significativement entre les hommes et les femmes. Cela

pourrait indiquer des différences dans le comportement de résiliation ou dans les besoins et préférences d'assurance entre les genres. De plus, la comparaison des profils de distribution par genre selon les âges de souscription montre que (i) les hommes investissent plus tôt que les femmes entre 35 et 50 ans et (ii) que les femmes investissent relativement moins que les hommes entre 57 ans et 70 ans, ce qui peut être une information importante pour l'assureur quant à la collecte nette à évaluer sur ces tranches d'âge.

- **Pics de Résiliation :** Les années où l'on observe des pics de résiliation pour un genre ou les deux pourraient correspondre à des événements spécifiques justifiant une analyse plus approfondie. Par exemple, une augmentation des résiliations pourrait coïncider avec des périodes de crise économique, des changements législatifs affectant les assurances, ou des ajustements dans les offres de produits d'assurance.

Analyse des variables Indicateurs Économiques et Financiers :

Les graphiques présentant les évolutions des indicateurs économiques et financiers étant difficilement lisibles, nous avons recensé ci-après les éléments d'analyse majeurs qui en ressortent :

- **CPI (Indice des Prix à la Consommation) :** L'évolution de l'indice CPI montre des fluctuations au fil du temps, ce qui indique des variations dans le niveau général des prix des biens et services consommés par les ménages. Ces variations peuvent être dues à des facteurs tels que les changements dans la politique monétaire, les conditions économiques globales, ou des événements spécifiques affectant l'offre et la demande. Une tendance à la hausse du CPI peut indiquer une inflation, tandis qu'une tendance à la baisse peut signaler une déflation ou une période de stabilité des prix.
- **EUidx (Indice de l'Union Européenne) :** : L'évolution de l'indice EUidx montre l'évolution de l'incertitude économique au sein de l'Union Européenne. Des pics dans cet indice peuvent refléter des périodes d'incertitude accrue, souvent liées à des crises économiques, des tensions politiques, ou des événements géopolitiques majeurs affectant l'Union Européenne. Une diminution de l'indice suggère une période de stabilisation et une réduction de l'incertitude économique, ce qui peut être favorable à l'investissement et à la croissance économique.

Ces indicateurs, en étant analysés conjointement, offrent une vue d'ensemble sur l'état de l'économie et les perspectives économiques. Par exemple, une hausse du CPI en conjonction avec une augmentation de l'EUidx pourrait indiquer une période d'inflation dans un contexte d'incertitude économique élevée, nécessitant une attention particulière de la part des décideurs politiques et des investisseurs. Inversement, une stabilité ou une baisse du CPI accompagnée d'une réduction de l'EUidx pourrait signaler une période de stabilité économique et une confiance accrue dans l'économie.

L'évolution des taux d'intérêt à 1 an (rate1Y), 2 ans (rate2Y), et 10 ans (rate10Y) révèle plusieurs indications concernant les conditions de financement et les attentes du marché sur la période analysée :

- Variabilité des Taux : une variabilité des taux d'intérêt est constatée à court, moyen et long terme. Cette variabilité peut être influencée par divers facteurs tels que les politiques monétaires des banques centrales, les attentes d'inflation, et les conditions économiques globales.
- Tendances des Taux à Court vs Long Terme : La comparaison des taux à 1 an, 2 ans, et 10 ans met en lumière des moments où les taux à court terme sont plus élevés que les taux à long terme, ce qui peut indiquer une inversion de la courbe des taux. Une telle inversion est souvent interprétée comme un signe avant-coureur de récession. À l'inverse, lorsque les taux à long terme sont nettement plus élevés, cela peut refléter une anticipation de croissance économique et d'inflation.
- Réactions aux Événements Économiques et Politiques : Les fluctuations des taux d'intérêt peuvent également refléter la réaction des marchés à des événements économiques et politiques spécifiques. Par exemple, une baisse soudaine des taux peut être la conséquence d'une action de politique monétaire visant à stimuler l'économie, tandis qu'une hausse peut être due à une anticipation d'inflation ou à une amélioration des perspectives économiques.
- Comparaison des Échéances : La différence entre les taux à différentes échéances (1 an, 2 ans, 10 ans) offre un aperçu de la structure par terme des taux d'intérêt. Cette structure peut donner des indications sur les attentes du marché concernant l'évolution future des taux d'intérêt et de l'économie en général.

En résumé, l'analyse des taux d'intérêt à différentes échéances permet de mieux comprendre les attentes du marché, les politiques monétaires, et les perspectives économiques. Les variations observées dans les graphiques soulignent l'importance de surveiller ces indicateurs pour anticiper les tendances économiques et prendre des décisions éclairées en matière d'investissement et de financement.

Analyse des variables indicateurs économiques et sociétaux :

L'évolution des indicateurs unemploy (taux de chômage), industry (indice industriel), et RTV (indice de la vente au détail) fournissent des éléments intéressants en lien avec l'état de l'économie et l'aspect sociétal :

- Taux de Chômage (unemploy) : Les fluctuations du taux de chômage reflètent les changements dans la santé de l'économie. Une augmentation du taux de chômage peut indiquer une contraction économique, où les entreprises réduisent leurs effectifs en réponse à une baisse de la demande. Inversement, une diminution du taux de chômage suggère une expansion économique, avec des créations d'emplois suivant une augmentation de la demande. L'évolution

de l’indice montre des périodes de stabilité économique alternant avec des périodes de volatilité, reflétant les cycles économiques.

- Indice Industriel (industry) : L'évolution de l'indice industriel offre un aperçu de la performance du secteur industriel, qui est un moteur important de la croissance économique. Une tendance à la hausse dans cet indice peut signaler une augmentation de la production industrielle, souvent associée à une demande accrue et à une économie en croissance. Une tendance à la baisse peut indiquer une réduction de la production, potentiellement due à une baisse de la demande ou à des défis économiques.
- Indice de la Vente au Détail (RTV) : Cet indicateur mesure la performance du secteur de la vente au détail, reflétant la confiance et le pouvoir d'achat des consommateurs. Une augmentation de l'indice RTV suggère que les consommateurs dépensent plus, ce qui est souvent un signe de confiance dans l'économie. Une baisse peut indiquer une réticence à dépenser, possiblement due à une incertitude économique ou à une baisse du revenu disponible.

En combinant ces trois indicateurs, on peut obtenir une vue d'ensemble de l'état économique et sociétal. Par exemple, une période où le taux de chômage diminue, tandis que les indices industriels et de vente au détail augmentent, peut indiquer une phase de reprise ou de croissance économique. À l'inverse, une augmentation du taux de chômage accompagnée d'une baisse des deux autres indices peut signaler une contraction économique.

Ces indicateurs sont importants pour les assureurs et les investisseurs car ils fournissent des informations clés sur les tendances économiques, permettant de prendre des décisions éclairées en matière de stratégie d’investissement et d’arbitrage.

C. Statistiques descriptives : analyse multivariée :

L’analyse multivariée peut mettre en lumière des interactions complexes entre l'âge lors de la souscription, le montant de couverture, et la résiliation des polices. Elle peut aider à identifier des modèles ou des tendances qui pourraient être utilisés par les assureurs pour affiner les stratégies de tarification, de marketing, ou de gestion des risques.

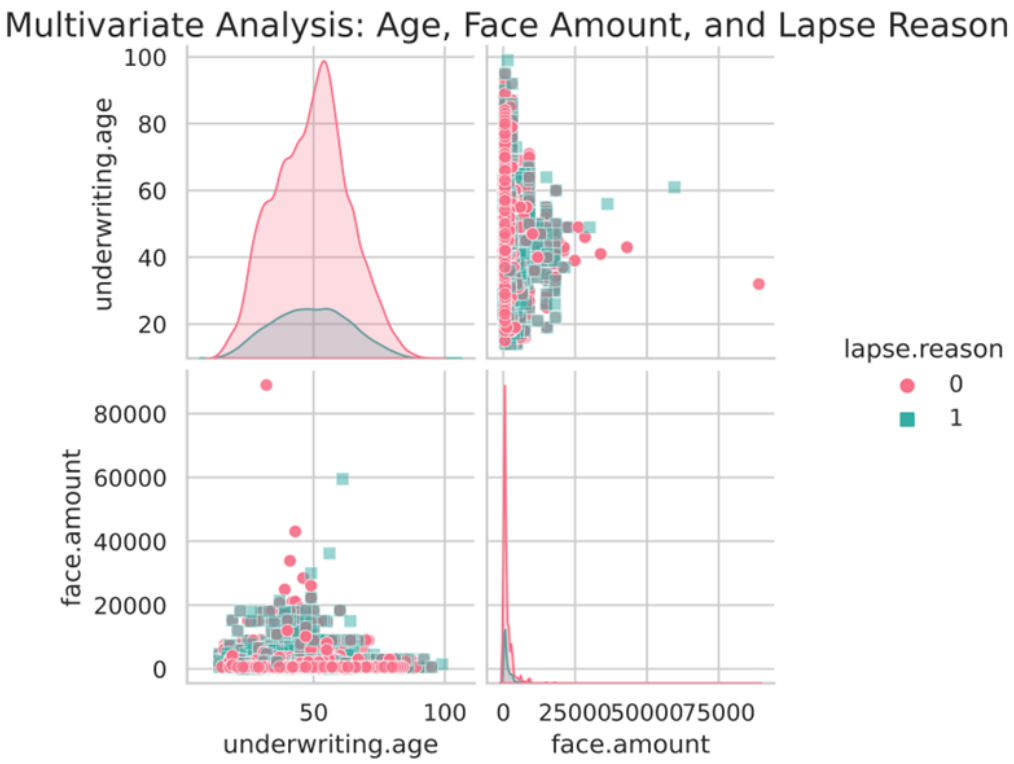


Figure : Analyse multivariée sur l'âge de souscription, le montant assuré et si résilié ou non

L'analyse multivariée des valeurs aberrantes (Multivariate Outlier Analysis) est une technique statistique utilisée pour identifier les observations qui se démarquent significativement des autres dans un ensemble de données multidimensionnel. Contrairement à l'analyse univariée, qui examine les valeurs aberrantes dans une seule variable à la fois, l'analyse multivariée prend en compte les relations entre plusieurs variables pour déterminer si une combinaison spécifique de valeurs est atypique.

- **Répartition de l'âge à la souscription et du montant assuré :** La distribution des points montre comment ces deux variables se comportent par rapport à la résiliation des polices. Il semble y avoir une large dispersion des âges lors de la souscription et des montants assurés, indiquant une variété dans les profils des souscripteurs. A noter que les montants assurés peuvent paraître faibles ; il s'agit ainsi ici très certainement des montants assurés issus des primes versées pendant cette fenêtre d'étude. Nous n'observons pas de valeurs aberrantes. Dans le cadre d'une optimisation de modèle, nous pourrions par la suite retirer les quelques observations qui semblent ne pas être représentatives de l'échantillon global (montants assurés supérieurs à 60K€ par exemple) afin d'éviter que le modèle établisse son apprentissage sur ces valeurs peu représentatives.
- **Relation entre l'âge de souscription et le montant assuré :** En examinant les graphiques, on peut observer s'il existe une tendance reliant l'âge lors de la souscription au montant de couverture. Par exemple, si les souscripteurs plus âgés tendent à avoir des montants de couverture plus élevés ou non. Ici, aucune tendance claire ne semble perceptible.

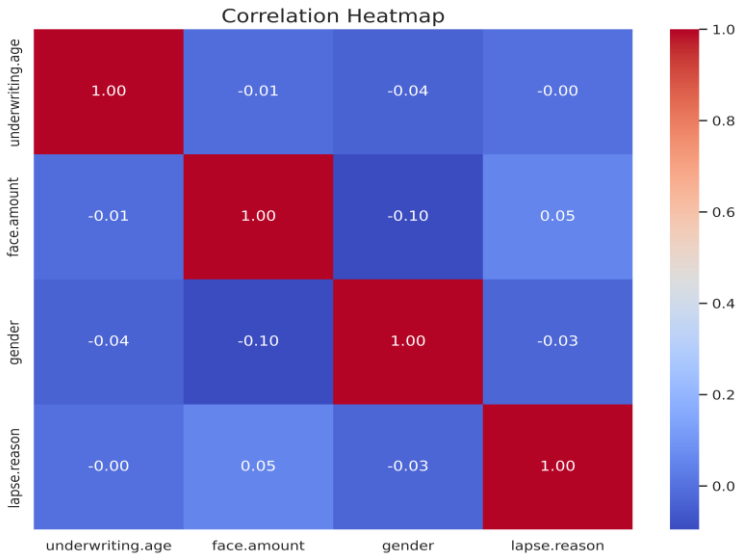


Figure : Heatmap de corrélation sur les critères de la clientèle

Cette heatmap de corrélation spécifiquement sur critères des assurés examine les relations entre l'âge lors de la souscription, le montant assuré, le genre, et l'état résilié ou non. Les valeurs de corrélation sont annotées sur la heatmap, fournissant une vue d'ensemble des relations linéaires potentielles entre ces variables.

- **Corrélations faibles à modérées :** Les valeurs de corrélation entre les variables semblent faibles à modérées. Cela indique qu'il n'y a pas de relations linéaires très fortes entre ces variables dans notre dataset.
- **Age lors de la souscription et autres variables :** L'âge lors de la souscription peut avoir une corrélation légère avec les autres variables, mais sans indication d'une forte relation linéaire. Cela suggère que l'âge seul ne détermine pas de manière significative le montant assuré, la probabilité de résiliation d'une police. D'autres variables seraient donc impliquées (comme les indicateurs économiques, financiers et sociaux).

La corrélation de Pearson, qui est utilisée par défaut dans les Heatmap, est inadaptée pour certaines variables catégorielles. En effet, Pearson mesure uniquement la relation linéaire entre deux variables continues et peut produire des résultats biaisés lorsqu'il est appliqué à des variables catégorielles comme 'Genre' ou 'Résiliation'. Il conviendrait de mesurer la corrélation entre deux variables catégorielles par le biais d'un test de Chi² et la corrélation entre une variable numérique et une variable catégorielle par le biais d'une analyse de la variance ANOVA (permet de voir si une variable catégorielle influence une variable numérique) :

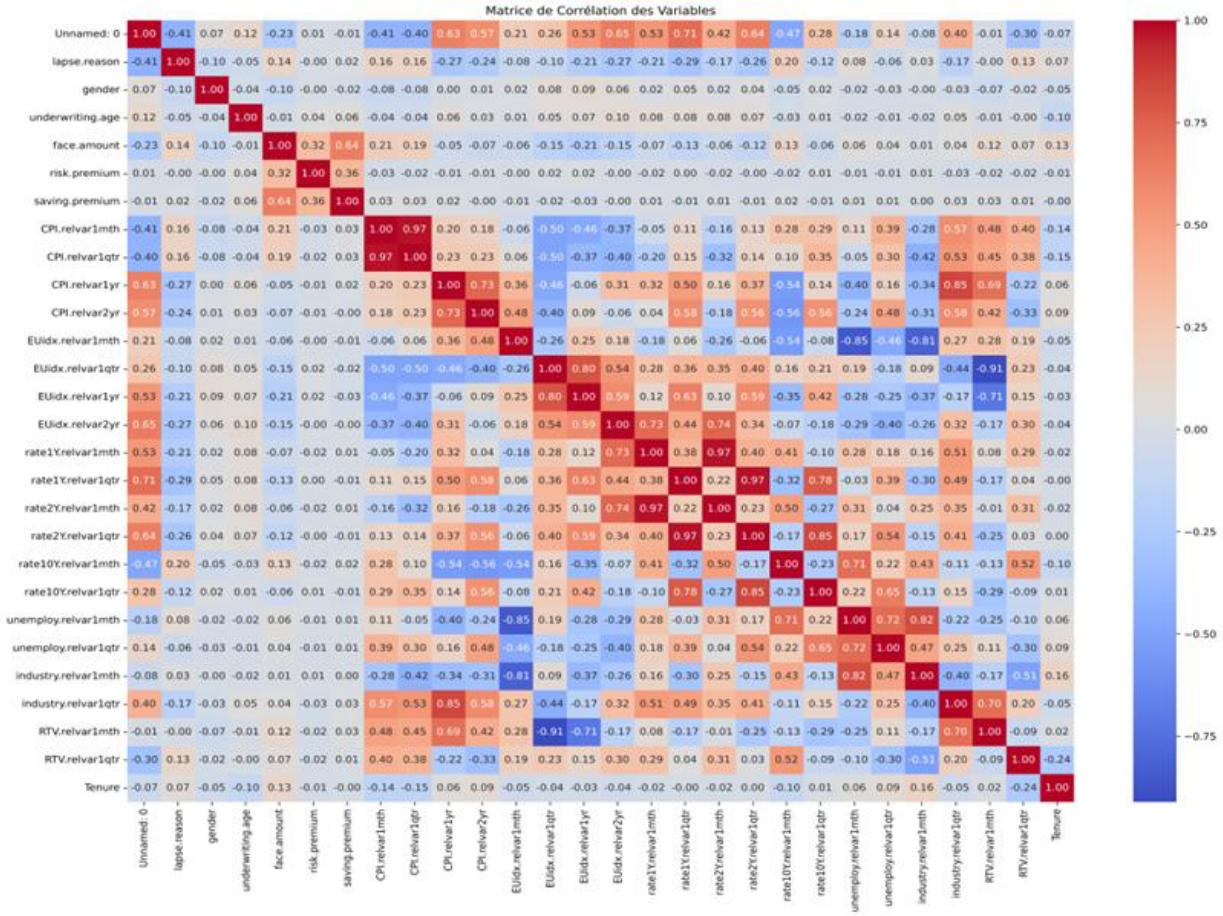
- **Montant assuré et risque de résiliation :** La corrélation entre le montant assuré et le risque de résiliation pourrait être intéressante à explorer davantage. Il pourrait être intéressant de

comprendre comment ces deux variables interagissent peut fournir des indications sur les motifs de résiliation en fonction du montant de couverture.

- **Genre et risque de résiliation :** L'analyse de la corrélation entre le genre et la raison de résiliation pourrait apporter un indicateur qui révèle des différences dans les motifs de résiliation entre les hommes et les femmes. Cela pourrait être dû à des facteurs sociodémographiques ou à des différences dans les comportements de souscription.

Les décisions de résiliation des polices et les montants assurés ne sont pas fortement influencées par un seul facteur comme l'âge ou le genre, mais résultent probablement d'une combinaison de plusieurs facteurs. Une analyse plus approfondie, en utilisant des modèles statistiques ou de machine learning plus complexes, pourrait aider à dévoiler des relations non linéaires ou des interactions entre variables qui ne sont pas immédiatement apparentes dans une analyse de corrélation linéaire.

Nous décidons de procéder à l'analyse multivariée en intégrant cette fois-ci l'ensemble des données disponibles, présentée ci-après à titre illustratif (échelle peu lisible mais analysé ci-après).



Cette illustration de heatmap donne des notions de corrélation entre les variables, bien que certains coefficients de corrélation entre variables catégorielles ne doivent pas être pris en compte pour les

raisons évoquées plus haut (coefficient de corrélation de Pearson). Nous pouvons classer les relations entre les indicateurs en trois catégories principales :

- **Corrélations positives fortes :** Les indicateurs qui montrent une corrélation positive forte (valeurs proches de 1) sont ceux qui évoluent dans le même sens. Cela signifie que lorsque l'un augmente, l'autre a tendance à augmenter également. Ces corrélations fortes sont souvent observées entre des indicateurs qui partagent des facteurs économiques ou sociaux communs.
- **Corrélations négatives fortes** : Les indicateurs avec des corrélations négatives (valeurs négatives) évoluent en sens opposé. Si l'un augmente, l'autre a tendance à diminuer. Ces corrélations peuvent révéler des relations de substitution ou des effets contracycliques entre les indicateurs.
- **Absence de corrélation ou corrélations faibles** : Les paires d'indicateurs présentant des corrélations faibles ou proches de zéro n'ont pas de relation linéaire claire. Cela ne signifie pas nécessairement qu'il n'y a pas de relation entre eux, mais plutôt que si relation il y a, elle n'est pas linéaire ou est influencée par d'autres facteurs.

Voici quelques exemples significatifs des **fortes corrélations positives** qui ressortent et leurs implications potentielles :

- **Corrélations entre les variations relatives du CPI (Consumer Price Index)** : Les variations du CPI sur différentes périodes (1 mois, 1 trimestre, 1 an, 2 ans) montrent des corrélations positives élevées entre elles. Cela suggère que les mouvements de l'inflation sont cohérents sur différentes périodes. Une inflation croissante sur le court terme est souvent suivie par une tendance similaire sur le moyen et long terme. Pour les compagnies d'assurance, cela peut signifier que les coûts futurs liés aux sinistres et aux prestations pourraient augmenter, influençant ainsi la tarification des primes et la gestion des réserves.
- **Corrélations entre les variations relatives des taux d'intérêt** : De même, les variations des taux d'intérêt sur différentes périodes (par exemple, les taux à 1 an, 2 ans, et 10 ans) montrent également des corrélations positives élevées. Cela indique que les mouvements des taux d'intérêt sont généralement alignés sur différentes échéances. Les taux d'intérêt influencent directement le rendement des investissements des compagnies d'assurance. Une hausse des taux peut améliorer les rendements sur les investissements, tandis qu'une baisse peut les réduire.
- **Corrélations entre les variations relatives des indices boursiers et économiques** : Les variations des indices boursiers (comme l'indice européen) et économiques (comme le volume du commerce de détail et l'indice industriel) sur différentes périodes montrent également des corrélations positives. Cela reflète la tendance générale de l'économie et des marchés financiers à évoluer de manière cohérente sur différentes périodes. Pour les compagnies d'assurance, cela peut signifier que les conditions économiques générales (croissance économique, santé du

secteur industriel, consommation des ménages) sont des indicateurs importants pour évaluer les risques et les opportunités d'investissement.

Ces corrélations positives fortes entre les variables économiques et financières soulignent l'importance pour les compagnies d'assurance de surveiller de près les indicateurs économiques et financiers. Ces indicateurs peuvent avoir un impact significatif sur les décisions de gestion des risques, de tarification des produits d'assurance, et de stratégie d'investissement.

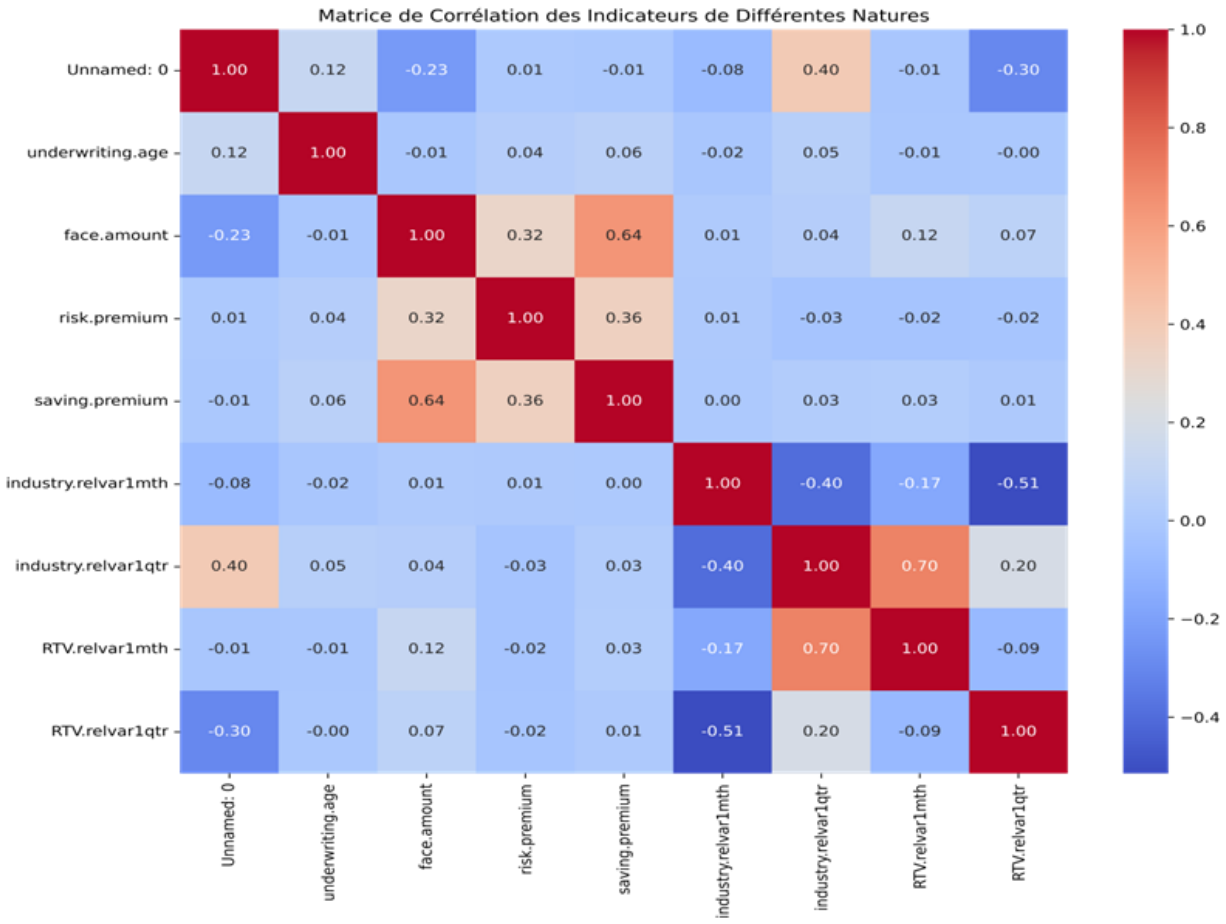
Par ailleurs, il conviendrait d'observer les corrélations entre les variables catégorielles et numériques qui nous intéressent par le biais d'une analyse de la variance ANOVA qui permet de voir si une variable catégorielle influence une variable numérique :

- **Corrélations entre les raisons de résiliation et les variables économiques** : Les corrélations entre les résiliations (lapse.reason) et plusieurs variables économiques telles que les variations relatives du CPI sur 1 an et 2 ans, ainsi qu'avec les variations relatives des taux d'intérêt sur 1 an et 2 ans pourraient indiquer que dans des périodes de volatilité économique ou de changements significatifs dans l'inflation ou les taux d'intérêt, les clients sont plus susceptibles de résilier leurs polices.
- **Corrélations entre l'âge de souscription et les variables économiques** : La corrélation entre l'âge de souscription (underwriting.age) et l'ancienneté de la relation avec le client (Tenure), pourrait suggérer que les polices souscrites à un âge plus avancé pourraient avoir une tendance à avoir une durée plus courte. Cela pourrait refléter une stratégie de souscription spécifique selon les moments de vie ou des besoins de couverture différents chez les souscripteurs plus âgés.

Ces corrélations offrent des indications sur la manière dont les conditions économiques et les caractéristiques des souscripteurs influencent les comportements de résiliation et la durée des polices. Pour les compagnies d'assurance, comprendre ces relations peut aider à mieux anticiper les risques de résiliation et à ajuster les stratégies de tarification et de gestion des polices en conséquence.

Nous avons révisé l'analyse en éliminant les redondances dues aux corrélations entre variables similaires mesurées sur différentes périodes afin qu'elle offre une perspective intéressante sur les relations entre les indicateurs économiques et financiers de natures différentes.

Ci-après le graphique Heatmap de corrélation révisé :



- **Variables économiques et financières** : Les indicateurs économiques tels que le taux de chômage, l'inflation (CPI), et les indices économiques peuvent être fortement corrélés avec des variables financières comme les taux d'intérêt. Par exemple, une augmentation de l'inflation (CPI) peut entraîner une hausse des taux d'intérêt pour contrôler cette inflation, reflétant une corrélation positive entre ces variables.
- **Indicateurs de marché et produits d'assurance :** Les variables liées aux produits d'assurance, telles que le montant assuré ou les primes d'épargne, peuvent être influencées par les conditions du marché économique et financier. Par exemple, dans un environnement de taux d'intérêt élevé, les produits d'épargne pourraient devenir plus attractifs, augmentant potentiellement les primes d'épargne. Ici ces variables n'ont pas l'air d'être fortement corrélées entre elles.
- **Variables démographiques et économiques :** L'âge de souscription (une variable démographique) pourrait avoir une corrélation avec des variables économiques ou financières en raison de comportements d'épargne ou d'investissement qui varient avec l'âge.

III. Présentation et construction d'un Modèle Prédictif

1. L'approche de l'apprentissage automatique (machine learning)

A. Concept général

L'apprentissage automatique, ou machine learning, est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données et de faire des prédictions ou de prendre des décisions sans être explicitement programmés pour chaque tâche.

Voici une explication étape par étape de la manière dont ce processus se déroule, spécifiquement dans le contexte de la prédiction des comportements de résiliation dans un portefeuille d'assurance vie.

A noter en premier lieu que dans le domaine du Machine Learning, la diversité des approches méthodologiques permet d'aborder la modélisation sous différents angles, en fonction de la nature des données et des objectifs spécifiques de l'analyse.

Ces méthodes se répartissent essentiellement en deux grandes catégories, chacune ayant ses particularités et domaines d'application :

a. Apprentissage Supervisé :

L'apprentissage supervisé est caractérisé par l'utilisation de données étiquetées pour l'entraînement des modèles. Dans ce contexte, les "étiquettes" correspondent aux résultats ou aux réponses que l'on souhaite prédire. Ces données d'entraînement comprennent donc à la fois les caractéristiques observées (les entrées) et les résultats associés (les sorties ou étiquettes). Le principal objectif de ces modèles est de minimiser l'erreur de prédiction, c'est-à-dire la différence entre les valeurs prédites par le modèle et les valeurs réelles (étiquettes).

L'apprentissage supervisé se prête idéalement à la prédiction des comportements de résiliation, où l'objectif est de modéliser la probabilité qu'un contrat soit racheté en fonction de divers facteurs explicatifs. Dans ce cadre, les données historiques du portefeuille, comprenant des informations sur les résiliations passées (étiquetées comme telles), servent à entraîner des modèles prédictifs. Ces modèles apprennent à associer des patterns spécifiques dans les données à la probabilité d'une résiliation.

- **Classification pour les sorties :**

Lorsque nous abordons la prédiction de résiliation comme un problème de classification, chaque contrat est étiqueté comme "résilié" ou "non résilié". Le modèle s'efforce alors de prédire cette catégorie en se

basant sur des caractéristiques telles que l'âge du souscripteur, la durée du contrat, les conditions économiques, etc.

- **Régression pour estimer la probabilité des sorties :**

Lorsque les étiquettes sont quantitatives, comme par exemple la prédiction du montant des résiliations en fonction des autres caractéristiques, on peut utiliser des modèles de régression. Le but est de prédire une valeur numérique continue selon les corrélations observées entre les variables (nuages d'observations). Si nous adoptons une approche de régression sur l'étiquette des résiliations, cela permettrait de quantifier la probabilité de résiliation ou le montant potentiel des sorties en fonction des corrélations entre les variables, offrant ainsi une perspective sur le risque de résiliation.

L'application de l'apprentissage supervisé dans un contexte transactionnel permet de modéliser directement la probabilité de résiliation en se basant sur des données historiques étiquetées. Cette approche est essentielle pour les acquéreurs potentiels qui cherchent à comprendre les niveaux de risque associés à chaque segment du portefeuille et à évaluer l'impact financier potentiel des résiliations futures.

b. Apprentissage Non Supervisé :

À l'opposé, l'apprentissage non supervisé ne s'appuie pas sur des données étiquetées. Les algorithmes explorent les données brutes pour y déceler des structures ou des motifs sans référence préalable. L'objectif ici est de regrouper les données en fonction de leurs similarités, sans connaître les catégories ou les résultats à l'avance. L'apprentissage non supervisé peut donc être utilisé pour segmenter le portefeuille en groupes homogènes de contrats ou de souscripteurs. Cette segmentation aide à identifier des profils de risque de résiliation ou des tendances comportementales sans se baser sur des étiquettes prédéfinies.

- **Clustering pour identifier les groupes homogènes de risque :**

Une application typique de l'apprentissage non supervisé est le clustering, ou segmentation, qui consiste à organiser les données en groupes (ou clusters) présentant des caractéristiques communes. Cela nous permet donc d'identifier les sous-groupes naturels au sein d'un ensemble de données, de repérer les comportements types ou les profils d'utilisateurs similaires. Le clustering permet de rassembler les groupes de contrats présentant des risques similaires de résiliations.

En résumé, le choix entre apprentissage supervisé et non supervisé dépend principalement de la nature des données disponibles et de l'objectif de l'analyse. Tandis que l'apprentissage supervisé est orienté vers la prédiction précise de valeurs ou de catégories spécifiques, l'apprentissage non supervisé vise à explorer et à comprendre la structure intrinsèque des données sans préjugés. La mesure des performances de prédiction de chacun des modèles pourra apporter du confort sur la fiabilité du choix du modèle pour la prédiction.

En intégrant ces deux approches d'apprentissage dans notre étude, nous pouvons non seulement prédire les comportements de résiliation avec une précision à évaluer selon les modèles mais aussi comprendre les structures sous-jacentes et les dynamiques au sein du portefeuille. Cela offre une base solide pour les assureurs afin d'élaborer des stratégies proactives de rétention, en alignant les échanges avec la clientèles et interventions de conseils personnalisés avec les profils de risque identifiés. In fine cela donne des indications précieuses pour l'assureur pour l'optimisation de ses décisions financières et opérationnelles au sein de l'entreprise.

Dans un contexte transactionnel, l'intégration des méthodes de machine learning pour évaluer le risque de résiliation apporte une dimension analytique avancée qui peut transformer l'approche de due diligence dans les transactions. Elle permet aux acquéreurs d'appréhender de manière plus fine et plus éclairée le risque inhérent au portefeuille d'assurance vie en UC qui fait l'objet de cession, facilitant ainsi la prise de décision stratégique et la négociation des termes de la transaction.

En définitive, l'utilisation judicieuse de ces techniques apporte une connaissance supplémentaire qui peut contribuer à sécuriser les investissements et à optimiser les stratégies de gestion post-acquisition.

B. Préparation et nettoyage des données : une étape fondamentale

La préparation et le nettoyage des données constituent la première et l'une des étapes les plus cruciales dans le processus d'apprentissage automatique, particulièrement dans notre étude axée sur la prédiction des comportements de résiliations dans un portefeuille d'assurance vie. Cette phase initiale est essentielle pour assurer l'intégrité et la qualité des analyses ultérieures, posant ainsi les fondations solides nécessaires à toute modélisation prédictive fiable.

Notre démarche a débuté par une pré traitement du dataset 'eusavingULnoPSperYr', issu de la collection CASdatasets. Nous avons d'abord éliminé les informations redondantes ou non pertinentes pour notre objectif d'analyse, telles que la première colonne numérotant simplement les lignes et la colonne de fréquence de paiement des primes, jugée non essentielle pour notre modèle. Cette simplification vise à éliminer le bruit et à concentrer notre attention sur les variables qui influencent directement les décisions de résiliation.

Ensuite, nous avons procédé à l'encodage des variables catégorielles, transformant des données qualitatives en données quantitatives pour permettre leur traitement par des algorithmes d'apprentissage automatique. Par exemple, le statut de résiliation du contrat (Surrender, Claim ou In force) a été codifié en valeurs binaires, facilitant ainsi l'identification des cas de résiliation. De même, la distinction de genre a été convertie en un format numérique, standardisant les données pour une analyse plus cohérente.

L'ajout du champ 'Tenure', représentant l'ancienneté de la relation client, illustre notre volonté de capter l'évolution temporelle des engagements des souscripteurs, enrichissant notre modèle de données dynamiques et pertinentes. Par ailleurs, la suppression de la colonne 'termination.date' a été décidée pour

éviter toute fuite d'information (data leakage), qui pourrait fausser la prédiction en introduisant des éléments trop évidents liés à l'événement de résiliation.

La transformation de 'issue.date' en deux variables distinctes, 'issue.year' et 'issue.month', nous permet de saisir plus finement l'impact temporel sur les comportements de résiliation, offrant une granularité accrue dans l'analyse des tendances et des cycles. En outre, nous avons introduit un calcul spécifique pour déterminer l'année d'observation de chaque ligne de données. Étant donné que les entrées du dataset se répètent pour une même police tant qu'elle reste en portefeuille, il était essentiel de déduire l'année d'observation correspondante à chaque entrée. Ce calcul nous aide à suivre l'évolution de chaque police au fil du temps, offrant une perspective dynamique sur la durée de vie et le comportement de résiliation de chaque contrat.

Cette étape de préparation et de nettoyage des données est impérative pour garantir la validité de notre étude. Elle influence directement la capacité des modèles d'apprentissage à identifier correctement les patterns et à générer des prédictions fiables. En épurant les données, en enrichissant le dataset avec des variables significatives, et en éliminant toute source potentielle de biais, nous posons les jalons d'une analyse robuste et éclairée, essentielle pour comprendre les dynamiques complexes à l'œuvre dans les décisions de résiliation des souscripteurs.

C. Analyse statistique approfondie dans le processus d'apprentissage automatique

L'analyse statistique approfondie constitue une étape cruciale dans le processus d'apprentissage automatique. Cette phase d'analyse va bien au-delà d'une simple inspection des données ; elle implique une exploration systématique et rigoureuse pour déceler les relations, les tendances et les anomalies qui peuvent influencer les décisions de résiliation.

- **Exploration et compréhension des données :** avant de construire des modèles prédictifs, il est essentiel de plonger dans les données pour en saisir la structure et les dynamiques sous-jacentes. Cette exploration commence par des statistiques descriptives, qui résument les caractéristiques principales des données à travers des mesures telles que la moyenne, la médiane, l'écart-type, et les quartiles. Ces mesures fournissent une vue d'ensemble de la distribution et de la centralité des données, permettant d'identifier les premières tendances et les valeurs atypiques.
- **Identification des corrélations et des patterns :** l'étape suivante consiste à examiner les relations entre différentes variables. En utilisant des techniques telles que la corrélation de Pearson pour les variables continues ou le test du Chi-deux (X^2) pour les variables catégorielles, les analystes peuvent détecter des associations significatives qui méritent une investigation plus approfondie. Cette analyse de corrélation aide à comprendre comment les différents facteurs interagissent entre eux et leur influence potentielle sur la probabilité de résiliation.
- **Visualisation des données :** la visualisation joue un rôle clé dans l'analyse statistique, offrant une représentation graphique des données qui peut révéler des découvertes non évidentes à

travers des tableaux ou des calculs statistiques. Des graphiques tels que les histogrammes, les boîtes à moustaches et les scatter plots permettent de visualiser la distribution des données, d'identifier les outliers et de comprendre les relations entre les variables.

- **Préparation pour la modélisation :** l'analyse statistique approfondie prépare le terrain pour la modélisation en fournissant une compréhension solide des données. Cette étape assure que les modèles d'apprentissage automatique seront construits sur une base informative et pertinente, augmentant ainsi leur capacité à générer des prédictions précises et fiables. En identifiant les variables significatives et en comprenant leur interaction, les analystes peuvent concevoir des modèles plus sophistiqués et adaptés aux spécificités du comportement de résiliation dans les assurances vie en UC.

L'analyse statistique approfondie est une étape indispensable qui enrichit la qualité de l'analyse prédictive en machine learning. Elle permet non seulement de valider les hypothèses initiales mais aussi de découvrir de nouveaux patterns et relations, essentiels pour élaborer des stratégies de gestion des résiliations basées sur des données probantes.

D. Construction des modèles de Machine learning (apprentissage, validation et sélection)

Le déploiement des algorithmes de machine learning est une étape cruciale pour permettre aux modèles en question d'apprendre et de prédire la variable cible. Cette phase implique l'entraînement, la validation des modèles et la sélection du modèle de machine learning le plus pertinent capable d'apprendre à partir des données et de faire des prédictions précises sur les comportements futurs de résiliation.

a. Les différents modèles de machine learning

Dans le cadre de notre étude visant à estimer le risque de résiliation dans un portefeuille d'assurance vie, nous allons déployer une série de modèles de machine learning, chacun ayant ses forces et particularités. L'objectif par la suite sera de comparer leur efficacité et de choisir le plus adapté pour notre contexte spécifique. Voici une explication détaillée de chaque modèle sélectionné :

- **Régression Logistique pour réponse binaire :** la Régression Logistique est particulièrement adaptée aux problèmes de classification binaire, comme la prédiction de la résiliation (oui ou non) d'un contrat d'assurance. Ce modèle estime la probabilité qu'un événement se produise en fonction de plusieurs variables explicatives, offrant une interprétation directe et intuitive des facteurs influençant les décisions de résiliation.
- **Modèle des arbres de décision (Decision Trees) :** les arbres de décision segmentent l'espace des données en un ensemble de règles simples basées sur les valeurs des variables explicatives. Chaque nœud de l'arbre représente une décision basée sur une seule variable, et chaque feuille représente une prédiction. Bien que simples à comprendre et à interpréter, les arbres de décision

peuvent être sujets au surajustement, surtout lorsqu'ils sont très profonds (longueur de la chaîne de décision entre la *root node* jusqu'à la plus longue *leaf node*).

- **Modèle des Forêts Aléatoires (Random Forest) :** les Forêts Aléatoires améliorent les arbres de décision en créant un ensemble de nombreux arbres, chacun entraîné sur un sous-ensemble aléatoire des données et des variables. La prédiction finale est obtenue par un vote majoritaire ou une moyenne des prédictions de tous les arbres. Ce modèle est réputé pour sa robustesse et sa capacité à gérer des données de haute dimensionnalité sans surajustement significatif.
- **XGBoost (eXtreme Gradient Boosting) :** XGBoost est une implémentation optimisée de l'algorithme de boosting de gradient, qui construit séquentiellement un ensemble de modèles faibles (généralement des arbres) de manière à ce que chaque nouveau modèle corrige les erreurs du modèle précédent. XGBoost est particulièrement apprécié pour sa performance et son efficacité. Il est d'ailleurs souvent utilisé dans les compétitions de data science pour sa capacité à produire des résultats de haute précision.
- **Modèle des Réseaux de Neurones :** les Réseaux de Neurones sont des modèles inspirés du fonctionnement du cerveau humain, capables de capturer des relations complexes et non linéaires entre les variables. Ils sont composés de couches de neurones avec des connexions pondérées entre eux. Les Réseaux de Neurones nécessitent généralement un grand volume de données pour l'entraînement et sont plus complexes à interpréter, mais ils peuvent capturer des patterns subtiles et complexes non détectables par d'autres modèles.

Nous verrons que théoriquement la régression pénalisée ne semble pas forcément être le modèle le plus adéquat pour notre problématique de classification. Le modèle XGBoost est très souvent le plus performant mais il arrive parfois que d'autres modèle le surpasse, ce qui explique que nous procéderons à l'entrainement d'une série de modèles avant de choisir le plus pertinent d'entre eux.

En intégrant et en comparant ces différents modèles, nous pouvons tenter d'évaluer le risque de résiliation dans le portefeuille d'assurance vie à disposition, en tenant compte de la diversité et de la complexité des comportements de résiliation.

b. Entraînement des Modèles (apprentissage)

L'entraînement des modèles consiste à les ajuster sur les données historiques pour qu'ils apprennent à identifier les signaux prédictifs de résiliation. Cette phase nécessite une division rigoureuse des données en ensembles d'entraînement et de test pour évaluer objectivement la performance du modèle.

La séparation du dataset étudié en ensembles d'entraînement et de test a été réalisé selon les dimensions des ensembles ci-après en veillant à bien garder le même niveau de représentation des états résiliés et non résiliés (et des genres, de l'âge de souscription moyen, etc.) :

- Ensemble d'entraînement : (34 994 observations avec 17% de résiliation)

- Ensemble de test : (8749 observations avec 17% de résiliation)

Pour rappel, 41% des polices ont résiliés pendant la période d'étude mais nous cherchons les évènements qui déclenchent les résiliations, ainsi nous avons décidé de regarder chaque ligne distinctivement représentant chacune des années d'observation pour chaque police pour observer les évènements déclencheurs de la résiliation. Ainsi 17% des lignes du datatset affichent le statut résilié. Environ 62% des souscripteurs sont des hommes, contre 38% de femmes. L'âge moyen lors de la souscription est d'environ 49 ans.

Nous avons ainsi 34 994 observations dans l'ensemble d'entraînement et 8 749 observations dans l'ensemble de test, chacun avec 27 caractéristiques après avoir retiré les colonnes spécifiées dans la préparation de la base de données avant entrainement.

c. Validation, optimisation et sélection du modèle le plus pertinent

Une fois les modèles entraînés, ils doivent être validés en évaluant leur erreur de prédiction et en s'assurant qu'ils ne souffrent ni de surajustement ni de sous-ajustement. Les modèles peuvent être optimisés pour maximiser leur performance en ajustant les paramètres du modèle et en ajustant la base de données d'entrée (retraits de certaines occurrences qui génèrent un sur ou sous ajustement, ou qui peut biaiser l'apprentissage).

Il est possible d'utiliser la validation croisée pour évaluer la capacité généralisatrice du modèle, en testant sur plusieurs sous-ensembles des données pour s'assurer de sa stabilité et de sa fiabilité.

Plusieurs métriques peuvent être utilisées pour évaluer les performances des modèles entrainés, comme l'aire sous la courbe ROC (AUC-ROC), le R^2 pour la régression, les matrices de confusion ou encore d'autres scoring de performance précisés ci-après.

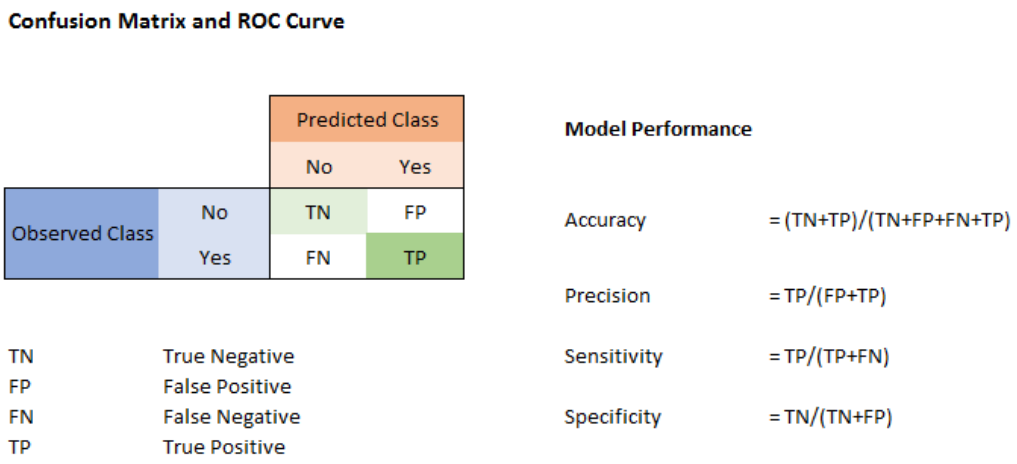


Figure : Matrice de confusion et autres métriques de performance des modèles

Interprétation des Matrices de Confusion

Chaque matrice de confusion montre le nombre de vrais positifs (TP), faux positifs (FP), vrais négatifs (TN) et faux négatifs (FN) pour le modèle correspondant. Ces valeurs permettent d'évaluer plus précisément la performance de chaque modèle en termes de classification.

- Vrais Positifs (*TP*) : Les instances correctement prédites comme positives.
- Faux Positifs (*FP*) : Les instances incorrectement prédites comme positives.
- Vrais Négatifs (*TN*) : Les instances correctement prédites comme négatives.
- Faux Négatifs (*FN*) : Les instances incorrectement prédites comme négatives.

Les modèles avec un nombre élevé de *TP* et *TN* et un nombre faible de *FP* et *FN* sont considérés comme ayant de meilleures performances.

Les matrices de confusion fournissent une vue détaillée permettant de comprendre comment chaque modèle performe en termes de classification. Ces informations peuvent être complétées de l'analyse des scores de performance suivants :

- **La précision** mesure la proportion des identifications positives qui sont effectivement correctes. Elle est calculée en divisant le nombre de vrais positifs par le nombre total de positifs prédits (la somme des vrais positifs et des faux positifs). Une précision élevée indique que lorsque le modèle prédit une classe positive, il est très probable que cette prédiction soit correcte :

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

- **Le rappel (recall)** mesure la proportion des vrais positifs qui sont correctement identifiés par le modèle. Il est calculé en divisant le nombre de vrais positifs par le nombre total de positifs réels (la somme des vrais positifs et des faux négatifs). Un rappel élevé signifie que le modèle est capable d'identifier une grande partie des cas positifs réels.

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

- **Le score F1** est une mesure qui combine la précision et le rappel en une seule métrique, offrant un équilibre entre les deux. Il s'agit de la moyenne harmonique entre ces deux métriques. Un score F1 élevé indique que le modèle est performant à la fois en termes de précision et de rappel.

$$Score\ F1 = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$$

- **Le score AUC (aire sous la courbe ROC)** évalue la capacité du modèle à distinguer entre les classes. L'AUC score mesure la performance globale du modèle à travers tous les seuils de classification. Un score AUC proche de 1 indique une excellente performance du modèle. Bien que l'AUC lui-même ne soit pas calculé directement à partir des TP, FP, TN et FN, la courbe ROC est construite en utilisant le taux de vrais positifs (*TPR*, un autre nom pour le rappel) et le taux de faux positifs (*FPR*), qui sont dérivés de ces valeurs.

$$TPR\ (Recall) = \frac{TP}{TP + FN} \quad ; \quad FPR = \frac{FP}{FP + TN}$$

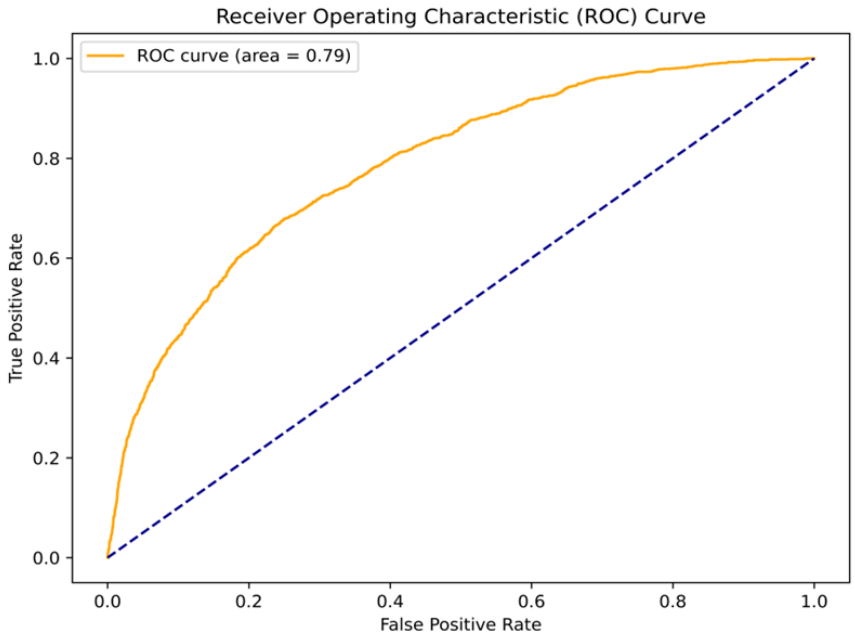


Figure : Illustration d'une courbe ROC (79% de AUC ici)

Interprétation des résultats :

Ces métriques de performance fournissent une vue d'ensemble de l'efficacité du modèle à différents niveaux et sont essentielles pour évaluer et comparer les performances des modèles. Il est crucial d'interpréter les résultats du modèle de manière contextualisée, en analysant comment les variables contribuent aux prédictions et en évaluant la pertinence des prédictions.

E. Intégration des résultats dans le contexte de l'étude

Les résultats obtenus grâce à l'analyse doivent être intégrés dans la stratégie transactionnelle. Cela implique de traduire les prédictions de risque de résiliation en estimations financières, en considérant l'impact sur la valeur du portefeuille.

Cette démarche permet d'apporter une connaissance supplémentaire côté acheteur dans le processus transactionnel permettant une évaluation éclairée de la valeur du portefeuille d'assurance vie en UC et facilitant les décisions stratégiques dans le cadre de la transaction.

F. Récapitulatif des étapes des méthodes d'apprentissage automatique

Le schéma suivant récapitule les différentes étapes de prédiction par les méthodes de machine learning tout en précisant les réitérations possibles pour une meilleure optimisation des modèles :

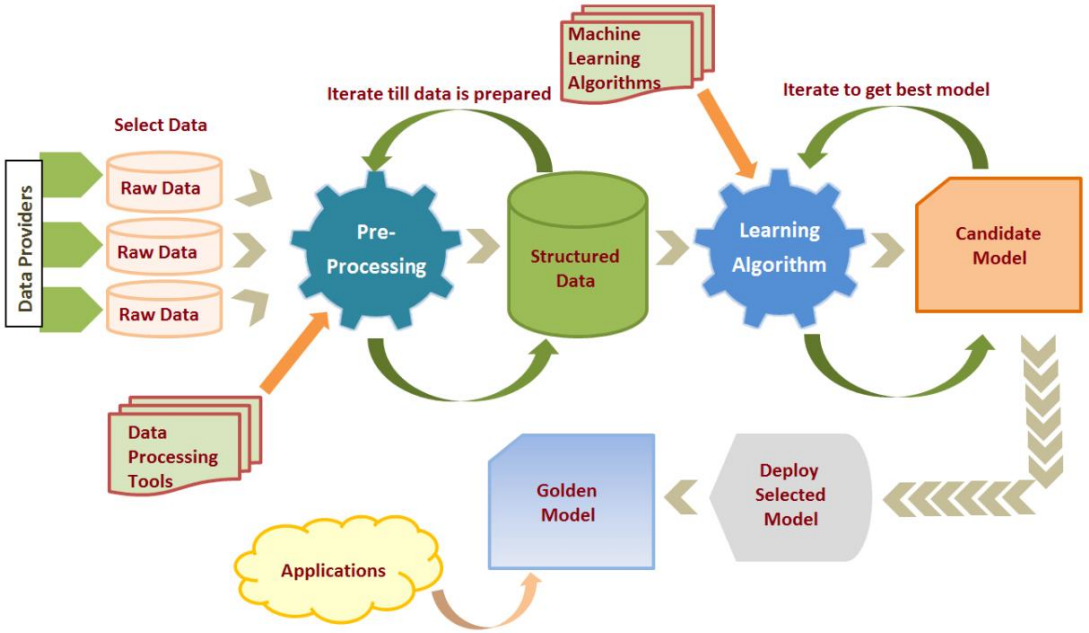


Figure : Récapitulatif des étapes des méthodes d'apprentissage automatique

2. Présentation théorique des modèles

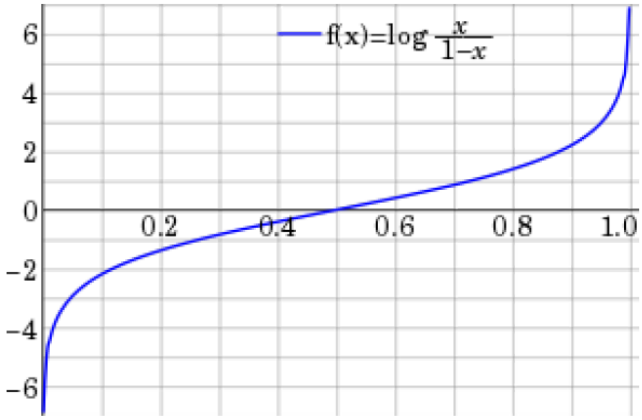
Nous nous sommes inspirés du mémoire de CHOUKARAH, R. (2019) sur la « Prédiction des rachats sur le portefeuille épargne AXA Japon) », dans le cadre de cette présentation théorique des modèles, et avons tenté d'apporter des compléments si cela est pertinent dans le cadre de notre problématique.

A. Régression Logistique pour réponse binaire

La Régression Logistique est un modèle de la famille des Generalized Linear Models (GLM). Ce modèle est particulièrement utile pour les problèmes de classification où les observations doivent être assignées à l'une des deux catégories possibles. La Régression Logistique établit un lien entre un événement et une combinaison linéaire de variables explicatives. Tandis que la régression linéaire suppose une variable dépendante suivant une distribution normale, la Régression Logistique à réponse binaire est un modèle statistique utilisé pour prédire l'issue d'une variable dépendante binaire de nature dichotomique (0 ou 1) à partir d'une ou plusieurs variables indépendantes. Ce modèle semble adapté à des situations où l'issue est binaire comme dans notre cas de figure, où l'on cherche à prédire les résiliations de contrats d'assurance.

La Régression Logistique consiste à modéliser la probabilité d'un événement cible binaire (la résiliation) en fonction des variables explicatives disponibles et s'opère en transformant une combinaison linéaire des prédicteurs en une probabilité comprise entre 0 et 1 grâce à la fonction logistique.

La Régression Logistique modèle la probabilité $P(Y = 1 \mid X)$ que la variable dépendante Y prenne la valeur 1 étant donné un ensemble de variables indépendantes X. La probabilité est modelée en utilisant la fonction logistique (ou Logit), qui transforme la combinaison linéaire des variables explicatives en une probabilité comprise entre 0 et 1.



Graphe du LOGIT

La **fonction Logit** est le cœur de la Régression Logistique. Elle est définie comme le logarithme naturel du rapport de cotes (odds ratio) de p par rapport à $1 - p$, où p est la probabilité de l'événement d'intérêt (par exemple, la résiliation d'un contrat). La fonction Logit s'écrit comme suit :

$$\text{Logit}(p) = \ln(\frac{p}{1-p})$$

La Régression Logistique assume que les log-odds (i.e. le logarithme naturel du rapport de probabilités (odds) qu'un événement se produise par rapport à sa non-occurrence) de la probabilité p peuvent être exprimés comme une combinaison linéaire des variables indépendantes. Cela conduit à l'équation suivante :

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Où $\beta_0, \beta_1, \ldots, \beta_n$ sont les paramètres du modèle à estimer.

Pour retrouver la probabilité p à partir des log-odds, nous utilisons la transformation inverse de la fonction Logit, qui est la fonction logistique :

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}$$

Cette fonction assure que la probabilité prédite reste toujours entre 0 et 1, quelles que soient les valeurs des variables explicatives.

Les coefficients β du modèle représentent l'effet log-odds d'une augmentation unitaire de la variable correspondante sur la probabilité de l'événement d'intérêt, tout en maintenant constantes les autres variables. Un β positif indique une augmentation de la probabilité de l'événement avec l'augmentation de la variable, tandis qu'un β négatif indique une diminution.

Les paramètres β du modèle sont estimés en maximisant la fonction de vraisemblance, ce qui revient à trouver les valeurs des paramètres qui rendent les données observées les plus probables. Cette méthode d'estimation est connue sous le nom de maximum de vraisemblance. Cette méthode cherche à trouver les valeurs des coefficients qui maximisent la probabilité (vraisemblance) des données observées.

La fonction de vraisemblance pour un modèle de Régression Logistique, donnant la probabilité des observations Y étant donné les prédicteurs X et les paramètres β, est définie comme :

$$L(\beta \mid X, Y) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

où :

- n est le nombre total d'observations,
- y_i est la valeur de la variable réponse pour l'observation i (0 ou 1),

p_i est la probabilité prédite que $Y = 1$ pour l'observation i, donnée par la fonction logistique :

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}$$

Pour faciliter la maximisation, nous utilisons le logarithme naturel de la fonction de vraisemblance, connu sous le nom de log-vraisemblance :

$$\ln(L(\beta \mid X, Y)) = \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)]$$

Pour maximiser la log-vraisemblance par rapport aux paramètres β, nous prenons sa dérivée par rapport à chaque β_j et égalons à zéro. Pour un paramètre β_j, la dérivée est :

$$\frac{\partial \ln(L(\beta \mid X, Y))}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij}(y_i - p_i)$$

où X_{ij} est la valeur de la j-ème variable explicative pour l'observation i.

Les équations obtenues par la dérivation de la log-vraisemblance ne peuvent généralement pas être résolues analytiquement pour les β_j. En pratique, pour obtenir une solution approximative satisfaisante de la maximisation mentionnée, les logiciels mettent en œuvre une procédure approchée. Cela justifie pourquoi les coefficients obtenus ne sont pas systématiquement identiques à chaque fois. Les différences observées dans les résultats peuvent s'expliquer par l'algorithme spécifique employé et le niveau de précision défini dans les paramètres de calcul.

Ainsi, des méthodes d'optimisation numérique telles que la méthode de Newton-Raphson ou des algorithmes de descente de gradient sont utilisées pour trouver les valeurs de β qui maximisent la log-vraisemblance.

La méthode de Newton-Raphson est une technique d'optimisation numérique utilisée pour trouver rapidement les racines d'une fonction ou les points où elle s'annule. Basée sur l'utilisation des dérivées premières pour estimer la pente, cette méthode ajuste itérativement une estimation initiale jusqu'à convergence vers la solution. Bien qu'efficace et rapide pour des fonctions bien comportées près de la solution, son succès dépend fortement de l'estimation initiale et de la nature de la fonction étudiée.

Les algorithmes de descente de gradient sont des méthodes d'optimisation qui cherchent à minimiser une fonction de coût en se déplaçant itérativement dans la direction de la pente la plus raide, opposée au gradient de la fonction. En ajustant les paramètres de la fonction étape par étape, ces algorithmes convergent progressivement vers le minimum local ou global, rendant cette approche largement utilisée dans l'apprentissage automatique pour l'entraînement des modèles.

Une fois les paramètres β estimés, chaque coefficient β_j représente le changement dans les log-odds de l'événement $Y = 1$ pour une augmentation unitaire de la variable X_j, toutes les autres variables restant constantes. Un coefficient positif indique un effet augmentant la probabilité de l'événement, tandis qu'un coefficient négatif indique un effet diminuant cette probabilité.

Adaptation du modèle pour notre jeu de données

Nous utiliserons en Python un modèle fondé sur la Régression Logistique, qui correspond bien à la nature des données à notre disposition. Afin d'optimiser l'efficacité du modèle, en intégrant les avantages des régressions pénalisées à ceux du modèle logistique, nous mettons en œuvre une Régression Logistique pénalisée. Ce modèle suit les principes de base de la Régression Logistique, mais il est amélioré par l'introduction d'un terme de pénalisation L1 ou L2 à la fonction de vraisemblance.

B. Modèle des Arbres de Décision : Decision Trees

Les arbres de décision sont des modèles prédictifs utilisés en apprentissage supervisé. Ils permettent de représenter graphiquement et de résoudre des problèmes de décision à travers une structure arborescente.

Les arbres de décision sont populaires en raison de leur simplicité d'interprétation, de leur applicabilité à la fois aux tâches de classification et de régression, et de leur capacité à gérer des données hétérogènes.

Ainsi, un arbre de décision est composé de :

- Noeud racine ou (*root node*) : l'accès à l'arbre se fait par ce nœud.
- Nœud internes : correspondent à des tests sur les attributs des données. Chaque nœud interne divise l'ensemble des données en sous-ensembles selon les résultats des tests.
- Branches : représentent les résultats des tests effectués dans les nœuds et guident vers le sous-ensemble suivant de données ou vers une décision finale.
- Feuilles ou nœuds terminaux, qui indiquent les prédictions finales de l'arbre, c'est-à-dire les classes pour les problèmes de classification ou les valeurs continues pour les régressions.

L'apprentissage d'un arbre de décision consiste à construire la structure de l'arbre de manière itérative par division récursive de l'ensemble de données. Le choix du meilleur attribut pour diviser les données à chaque étape se fait en fonction de critères tels que l'indice de Gini, l'entropie ou l'erreur de classification pour les tâches de classification, et la variance ou l'erreur quadratique moyenne pour les régressions :

- **Indice de Gini :** Cette mesure évalue le degré d'impureté (ou de diversité) d'un nœud dans un arbre de décision. Un indice de Gini de 0 indique une pureté parfaite, où toutes les observations dans le nœud appartiennent à une seule classe. Plus l'indice de Gini s'éloigne de 0, plus le nœud est impur.
- **Entropie :** L'entropie est une mesure de l'incertitude ou du désordre dans un ensemble de données. En contexte d'arbre de décision, elle sert à évaluer l'homogénéité d'un nœud. Une entropie faible (proche de 0) signifie que la majorité des observations dans le nœud sont de la même classe, tandis qu'une entropie élevée indique une répartition plus équilibrée entre différentes classes.
- **Erreur de classification :** C'est la proportion d'erreurs commises par le modèle lorsqu'il assigne des observations aux classes. Dans le contexte des arbres de décision, c'est le taux d'observations mal classifiées dans un nœud. Un taux d'erreur de 0 signifie que toutes les observations sont correctement classifiées.

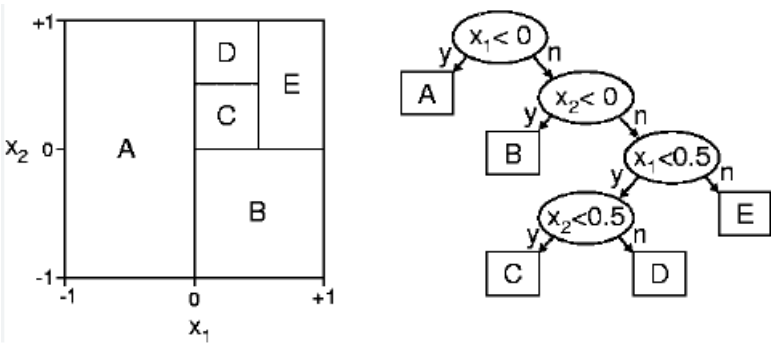
Ces mesures sont essentielles pour choisir le meilleur attribut lors de la construction d'un arbre de décision, car elles permettent d'identifier les divisions qui augmentent le plus l'homogénéité des nœuds, conduisant à un modèle plus précis.

Chaque nœud interne de l'arbre effectue un test sur l'une de ces caractéristiques, menant à différentes branches selon le résultat : pour les variables catégorielles, il y a une branche par valeur possible ; pour les variables numériques, les branches sont basées sur des intervalles de valeurs. Les feuilles de l'arbre représentent les différentes classes possibles. Pour classifier un nouvel élément, on le fait passer à travers l'arbre, de la racine vers une feuille, suivant les branches déterminées par les tests à chaque nœud, jusqu'à atteindre la classe prédite.

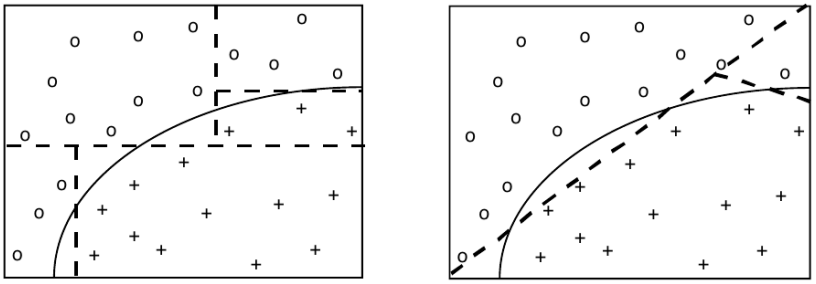
Construction de l'arbre

La construction d'un arbre de décision débute avec l'ensemble des données à la racine. On sélectionne l'attribut qui sépare le mieux les données selon le critère choisi, créant ainsi des branches vers des nœuds plus profonds. Cette procédure est répétée récursivement pour chaque sous-ensemble de données aux nœuds suivants, jusqu'à ce que l'une des conditions d'arrêt soit satisfaite, par exemple, que tous les éléments d'un sous-ensemble appartiennent à la même classe, qu'aucune amélioration significative ne soit possible, ou que l'arbre atteigne une profondeur maximale prédéfinie.

En essence, l'arbre divise de manière récursive l'ensemble des données, effectuant à chaque étape des tests sur les caractéristiques jusqu'à ce que chaque feuille de l'arbre représente un groupe d'éléments ayant la même valeur pour l'attribut cible, indiquant qu'ils appartiennent à la même classe. Les divisions récursives forment des régions distinctes dans l'espace des attributs où chaque région, correspondant à une feuille de l'arbre, est uniforme en termes de la variable cible.



Construction par partition récursive de l'espace



Séparation itérative de classes linéaires vs. non linéaires

À chaque étape de la construction d'un arbre de décision, l'objectif est de diviser le nœud actuel en deux sous-groupes aussi homogènes que possible. Cela peut se faire en utilisant une seule variable à la fois pour une séparation simple ou en combinant plusieurs variables pour une division plus complexe, ce qui peut simplifier l'arbre final mais rend les tests au niveau des nœuds plus complexes.

Les arbres de décision offrent l'avantage d'être intuitifs et faciles à comprendre, mais peuvent être sujets au surapprentissage si l'arbre devient trop complexe. Des techniques comme l'élagage sont alors utilisées pour simplifier l'arbre et améliorer sa généralisation à de nouvelles données.

Elagage des arbres

L'élagage des arbres est une technique cruciale pour éviter le surapprentissage, un phénomène où un modèle s'ajuste trop finement aux données d'apprentissage au détriment de sa capacité à généraliser à de nouvelles données. Un arbre avec des feuilles parfaitement homogènes peut parfaitement représenter l'échantillon d'apprentissage, mais risque de ne pas être performant sur de nouveaux échantillons. Ceci est dû au fait qu'il peut capturer du bruit ou des spécificités des données d'apprentissage qui ne sont pas représentatives de la population globale.

Dans la pratique, l'objectif est de construire un arbre aussi petit que possible tout en maintenant de bonnes performances de prédiction, en suivant le principe de parcimonie : à efficacité égale, un modèle plus simple est préférable car il est généralement plus stable et généralisable. Pour équilibrer la performance et la complexité, on évalue les modèles non seulement sur les données d'apprentissage, mais aussi sur un ou plusieurs échantillons de validation, des données étiquetées non utilisées dans la construction du modèle. Ce processus aide à sélectionner un modèle qui offre un bon compromis entre la précision des prédictions et la simplicité de la structure de l'arbre, améliorant ainsi sa capacité à être appliqué à de nouvelles données.

Le problème de surajustement

Le surajustement, ou overfitting, est un écueil courant en apprentissage automatique où un modèle s'adapte trop précisément aux données d'entraînement au point de perdre en généralisation. Pour évaluer ce phénomène, on utilise souvent un jeu de données de validation, traité comme un ensemble de test,

pour observer la stabilité de la performance du modèle sur des données non vues durant l'entraînement. Si un modèle performe bien sur l'échantillon d'apprentissage mais moins bien sur l'échantillon de validation, cela indique un surajustement.

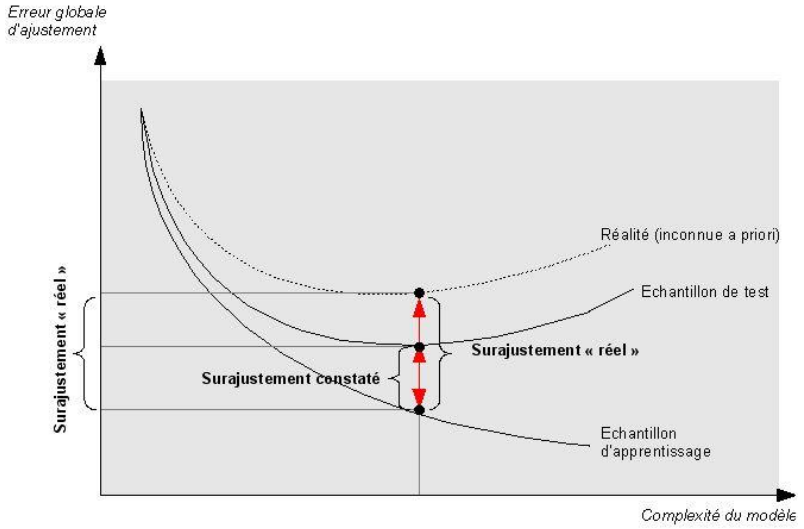


Illustration du surajustement

Un graphique ci-dessus illustre que l'erreur de prédiction par rapport au nombre de feuilles d'un arbre de décision (complexité de l'arbre) révèle souvent que, bien que l'erreur diminue sur l'ensemble d'apprentissage avec l'augmentation de la complexité, elle commence à augmenter sur l'ensemble de validation au-delà d'un certain point. Cela signifie que le modèle devient trop spécifique aux données d'entraînement et perd en capacité de prédiction sur de nouvelles données.

Pour contrer le surajustement, on utilise des techniques telles que le pré-élagage (arrêter la croissance de l'arbre avant qu'il devienne trop complexe) et le post-élagage (réduire la complexité d'un arbre après sa construction complète). Des cadres théoriques comme la Minimisation du Risque Structuré (Structured Risk Minimization - SRM) de Vapnik-Tchervonenkis, qui inclut la dimension VC (Vapnik-Chervonenkis), permettent de trouver un équilibre entre l'erreur sur l'ensemble d'apprentissage et l'erreur sur l'ensemble de test, aidant ainsi à construire des modèles à la fois précis et robustes. Ces approches algorithmiques sont complétées par des analyses comparatives de performance et de stabilité sur les échantillons d'apprentissage et de validation, pour optimiser la qualité et la généralisation du modèle.

Le pré-élagage

Le pré-élagage est une méthode employée pour prévenir le surajustement dans les modèles d'arbres de décision, en interrompant leur croissance avant qu'ils ne deviennent trop complexes. Cette technique implique de fixer des critères d'arrêt au cours de la construction de l'arbre, tels que la taille minimum d'un groupe de données dans un nœud ou un degré d'homogénéité déjà élevé qui rendrait inutiles des divisions supplémentaires. On peut aussi recourir à des tests statistiques pour déterminer si la division

d'un nœud apporte une amélioration significative dans la prédiction de la variable cible. Si les conditions du critère d'arrêt sont remplies, l'expansion de l'arbre s'arrête, évitant ainsi de produire un modèle excessivement complexe qui pourrait ne pas bien généraliser à de nouvelles données.

Le post-élagage

Le post-élagage est une technique qui permet de réduire la complexité des arbres de décision après leur construction complète. Cette approche se déroule en deux phases : initialement, on construit un arbre en cherchant à obtenir des feuilles aussi homogènes que possible, utilisant pour cela une partie de l'ensemble de données, souvent appelée échantillon d'apprentissage ou "growing set" pour éviter toute confusion. Ensuite, on procède à l'élagage de l'arbre en utilisant une portion différente de l'ensemble de données, connue sous le nom d'échantillon de validation ou de test, dans le but d'améliorer les performances globales du modèle. Cette phase d'élagage ajuste la structure de l'arbre pour éviter le surapprentissage, en supprimant les branches qui n'apportent pas d'amélioration significative à la prédiction sur de nouvelles données. Ce second ensemble utilisé spécifiquement pour l'élagage peut être appelé échantillon d'élagage ou "pruning set", distinguant clairement son rôle dans le processus de réduction de la complexité de l'arbre.

Les arbres de décision sont largement utilisés en apprentissage automatique pour leur facilité d'interprétation et leur flexibilité. Voici un aperçu de leurs principaux avantages et inconvénients :

Avantages :

- **Facilité d'interprétation et de visualisation :** Les arbres de décision peuvent être facilement compris et interprétés sans nécessiter de connaissances statistiques approfondies, ce qui les rend accessibles à un large public.
- **Gestion des données non linéaires :** Ils sont capables de capturer des relations non linéaires entre les variables sans nécessiter de transformation des données.
- **Traitement des variables catégorielles et continues :** Ils peuvent traiter à la fois des variables numériques et catégorielles sans nécessiter de prétraitement complexe.
- **Peu sensible aux valeurs aberrantes :** Les arbres de décision sont relativement robustes aux valeurs extrêmes dans les ensembles de données.
- **Aucune hypothèse sur la distribution des données :** Contrairement à d'autres méthodes statistiques, les arbres de décision ne font aucune hypothèse sur la distribution des variables dans l'ensemble de données.

Inconvénients :

- **Surapprentissage :** Sans élagage ou avec une profondeur de l'arbre non contrôlée, les arbres de décision sont sujets au surapprentissage, adaptant le modèle trop précisément aux données d'entraînement au détriment de la généralisation.

- **Stabilité :** De petits changements dans les données peuvent entraîner la génération d'arbres très différents, ce qui peut affecter la stabilité du modèle.
- **Tendance à la préférence pour les attributs avec plus de niveaux :** Les arbres de décision ont tendance à favoriser les variables avec un grand nombre de catégories, ce qui peut biaiser le modèle.
- **Problème de biais de classe :** Ils peuvent être biaisés envers les classes dominantes, ce qui peut nécessiter un équilibrage des classes pour de meilleurs résultats.
- **Optimalité :** Trouver l'arbre de décision optimal est un problème NP-complet, et les algorithmes heuristiques utilisés pour la construction d'arbres ne garantissent pas de trouver la meilleure solution possible.

Pour conclure, bien que les arbres de décision offrent de nombreux avantages en termes de facilité d'utilisation et de compréhension, ils présentent également des défis importants, notamment en termes de surapprentissage et de stabilité. Des techniques comme l'élagage des arbres et l'utilisation d'ensembles d'arbres (par exemple, les Forêts Aléatoires) sont des stratégies couramment utilisées pour surmonter certains de ces inconvénients.

C. Modèle des Forêts Aléatoires : Random Forest

Introduites par Breiman en 2001, les Forêts Aléatoires représentent une avancée significative dans les méthodes d'apprentissage automatique, offrant une alternative plus performante aux arbres de décision individuels, malgré une interprétabilité réduite. Cette méthode d'apprentissage en ensemble s'applique à la classification, la régression, et bien d'autres tâches, en construisant une multitude d'arbres de décision durant la phase d'entraînement. Pour la classification, la prédiction finale est déterminée par le vote majoritaire des classes prédites par chaque arbre, tandis que pour la régression, c'est la moyenne des prédictions qui est utilisée.

L'un des principaux avantages des Forêts Aléatoires est leur capacité à éviter le surajustement associé aux arbres de décision, grâce à la diversification des arbres entraînés sur des sous-ensembles de données. Breiman a étendu les travaux antérieurs en introduisant le concept de "bagging" (Bootstrap Aggregating), qui consiste à entraîner chaque arbre sur un échantillon aléatoire des données, avec remise. Cette technique, combinée à la sélection aléatoire des variables (feature selection) à chaque division dans les arbres, a été proposée indépendamment par Ho, Amit, et Geman, et permet de réduire la variance sans augmenter significativement le biais, ce qui conduit à des modèles plus robustes et généralisables.

L'algorithme de Random Forest effectue plusieurs étapes clés pour la construction et l'évaluation de ses arbres de décision.

Premièrement, il sélectionne aléatoirement, avec remise, des échantillons de la base de données d'entraînement pour chaque arbre, ce qui permet de créer des ensembles de données diversifiés.

Ensuite, lors de la construction de chaque arbre, un sous-ensemble aléatoire de variables est choisi à chaque division, ce qui augmente la diversité des arbres générés. Chaque arbre est construit jusqu'à sa taille maximale sans élagage, ce qui permet d'exploiter pleinement les informations disponibles. La prédiction finale est obtenue en agrégeant les prédictions de tous les arbres, par exemple, en prenant la moyenne pour une régression ou le vote majoritaire pour une classification.

Cet algorithme se fonde sur le principe du **bagging** (Bootstrap Aggregating), permettant de créer plusieurs arbres de décision indépendants et de les combiner pour obtenir une prédiction finale plus stable et fiable :

- **Bootstrap :** L'algorithme commence par générer plusieurs échantillons d'apprentissage à partir du jeu de données original, en utilisant le bootstrap. Le bootstrap est une méthode de rééchantillonnage avec remplacement, qui permet de créer plusieurs sous-ensembles de données de la même taille que l'ensemble original, mais avec certaines observations pouvant apparaître plusieurs fois dans chaque sous-ensemble.
- **Construction des arbres :** Pour chaque échantillon bootstrap généré, un arbre de décision est construit. Afin d'augmenter la diversité parmi les arbres, à chaque division d'un nœud dans l'arbre, une sélection aléatoire de variables (features) est effectuée parmi l'ensemble des variables disponibles, et la meilleure division est choisie uniquement parmi cet ensemble réduit. Cette étape est cruciale pour assurer que les arbres soient décorrélés et pour améliorer la performance globale du modèle.
- **Prédiction :** Une fois tous les arbres construits, la prédiction pour une nouvelle observation est obtenue en agrégeant les prédictions de tous les arbres. Pour une tâche de classification, cela se fait généralement par un vote majoritaire, où la classe la plus fréquemment prédite par les arbres est choisie. Pour une tâche de régression, la prédiction finale est souvent la moyenne des prédictions de tous les arbres.

Supposons un ensemble d'observations $X = \{x1, x2, \dots, xn\}$ avec les réponses correspondantes $Y = \{y1, y2, \dots, yn\}$. L'objectif est de construire un prédicteur $f(x)$ qui minimise l'erreur de prédiction, souvent mesurée par l'erreur quadratique moyenne (MSE) pour la régression :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$$

Pour appliquer le bagging, nous procédons comme suit :

- **Génération des sous-ensembles :** Générer B sous-ensembles bootstrap X_b* de la taille de X en effectuant un échantillonnage avec remplacement à partir de X.
- **Entraînement des modèles** : Pour chaque sous-ensemble bootstrap X_b*, entraîner un modèle f_b** sur ces données. Chaque modèle f_b* peut être un arbre de décision ou tout autre type de modèle d'apprentissage.
- **Agrégation des prédictions :** Pour une nouvelle observation x, calculer la prédiction agrégée en moyennant les prédictions de tous les modèles f_b* entraînés sur les sous-ensembles bootstrap.

Pour la régression, la prédiction agrégée est donnée par :

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{i=1}^{n} f_b*(x)$$

Pour la classification, la prédiction finale est le vote majoritaire parmi les prédictions de tous les modèles:

$$\hat{f}_{bag}(x) = \text{mode}\{f_1*(x), f_2*(x), \ldots, f_B*(x)\}$$

L'effet principal du bagging est la réduction de la variance de la prédiction finale, sans augmenter significativement le biais. Cette réduction de la variance est possible parce que, même si chaque modèle f_b* peut avoir une variance élevée, l'agrégation des prédictions de nombreux modèles entraîne une annulation mutuelle des erreurs individuelles, conduisant à une prédiction finale plus stable et plus précise. En d'autres termes, en moyennant les résultats de nombreux arbres, les erreurs individuelles de chaque arbre (qui peuvent être dues à la sur-adaptation aux données d'entraînement) se compensent mutuellement, résultant en une prédiction plus précise sur des données inédites.

Cependant, cette amélioration de la performance par le **bagging** s'accompagne d'une légère augmentation du biais et d'une réduction de l'interprétabilité du modèle, en comparaison avec un simple arbre de décision.

Ainsi, malgré ces compromis, les Forêts Aléatoires restent l'une des méthodes les plus efficaces et les plus utilisées en apprentissage supervisé, offrant un excellent équilibre entre précision, robustesse et capacité à gérer des jeux de données complexes.

Cette méthode offre plusieurs avantages. Elle est non seulement capable de gérer des données de grande dimensionnalité avec des interactions complexes entre les variables, mais elle est aussi robuste face aux données manquantes et réduit le risque de surajustement grâce à son approche d'ensemble. De plus, Random Forest peut être utilisé pour estimer l'importance des variables, offrant ainsi une interprétabilité utile pour comprendre les facteurs influençant les résiliations.

D. XGBoost : eXtreme Gradient Boosting

Dans le domaine du machine learning, il est souvent judicieux de consulter plusieurs sources plutôt que de dépendre d'une unique. Prédire un comportement ou estimer un montant numérique à l'aide d'un modèle constitue une base solide. Cependant, combiner les perspectives de milliers de modèles, chacun apportant son expertise sur certaines facettes des données, s'avère généralement plus efficace. C'est le principe des méthodes ensemblistes, parmi lesquelles le bagging et le boosting se distinguent. XGBoost, ou Extreme Gradient Boosting, appartient à cette catégorie d'approches. Nous explorons ici les principes fondamentaux et le mécanisme général sur lesquels repose XGBoost.

Divers algorithmes peuvent être déployés pour associer les caractéristiques des individus à une variable binaire représentant, par exemple, une catégorie. Chaque algorithme peut générer plusieurs modèles, chacun étant capable de distinguer les individus selon qu'ils appartiennent à une catégorie ou à une autre. Ces modèles sont souvent qualifiés de classifieurs faibles, c'est-à-dire qu'ils réalisent des classifications légèrement meilleures que le hasard, portant ainsi un minimum d'information utile.

Le boosting se distingue du bagging par la manière dont il pondère les différents classifieurs. Ainsi, lors de chaque prédiction, les classifieurs ayant effectué des prédictions justes se voient attribuer un poids plus important que ceux ayant failli. Adaboost, un des algorithmes de boosting, exploite cette idée en ajustant dynamiquement les poids pour accorder plus de valeur aux observations difficiles à prédire, valorisant ainsi les classifieurs performants là où d'autres échouent. Adaboost utilise des classifieurs préexistants et vise à leur attribuer des poids optimaux selon leur efficacité, et il a été adapté pour gérer des tâches de classification multi-classes.

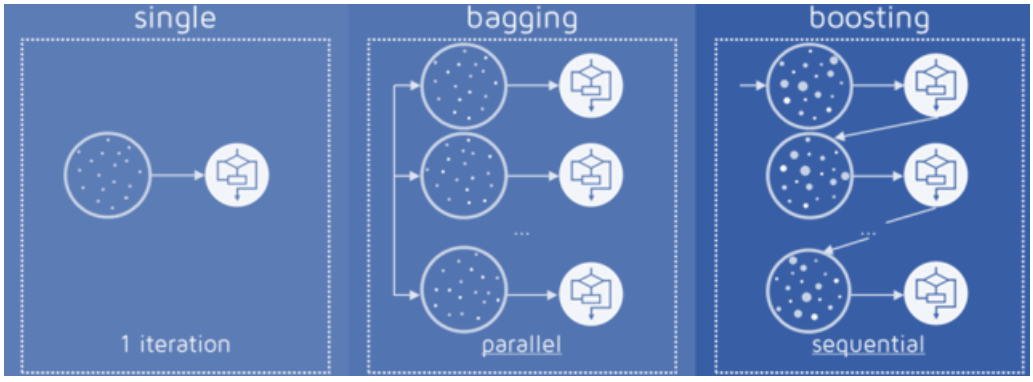


Figure : Single vs. Bagging vs. Boosting

La technique de boosting est fréquemment appliquée aux arbres de décision, où elle est connue sous le nom de Gradient Tree Boosting. Cette méthode vise à assembler plusieurs classifieurs de manière itérative, créant des "mini-classifieurs" qui sont souvent de simples fonctions paramétriques, typiquement des arbres de décision avec des critères de division spécifiques pour chaque branche. Le

classifieur composite final est alors le résultat de la combinaison pondérée de ces mini-classifieurs, à l'aide d'un vecteur de poids ω.

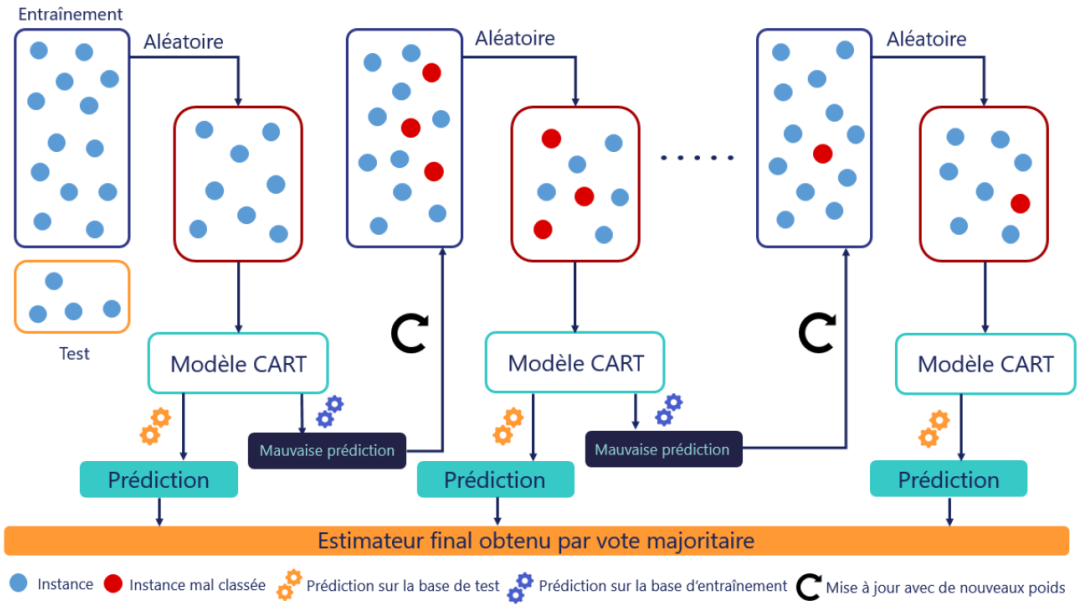


Figure : Fonctionnement de l'algorithme XGBoost

Pour élaborer ce classifieur composite, on suit plusieurs étapes :

- **Initialisation :** Sélectionner une combinaison arbitraire de mini-classifieurs avec des poids initiaux w_i et définir le classifieur composite.
- **Évaluation de l'erreur :** Calculer l'erreur générée par ce classifieur composite et identifier le mini-classifieur qui minimise cette erreur, ce qui équivaut à chercher dans l'espace des paramètres la fonction qui s'en approche le plus.
- **Optimisation :** Soustraire ce mini-classifieur du classifieur composite tout en ajustant son poids pour minimiser une fonction de perte donnée.
- **Itération :** Répéter ces étapes de manière itérative jusqu'à atteindre une performance satisfaisante.

Le classifieur obtenu à travers le gradient boosting est ainsi défini par les poids attribués à chaque mini-classifieur et les paramètres des fonctions choisies. Le processus revient à naviguer dans un espace de fonctions simples en effectuant une descente de gradient basée sur l'erreur, permettant d'affiner progressivement le modèle pour améliorer sa précision.

Le modèle final de gradient boosting est caractérisé par les poids attribués à chaque mini-classifieur et les paramètres des fonctions employées. Ce procédé consiste à naviguer dans un ensemble de fonctions simplifiées par une optimisation de gradient dirigée par l'erreur commise.

Essentiellement, cet algorithme élabore une série de modèles qui sont ensuite combinés à travers une moyenne pondérée de leurs prédictions ou par un système de vote majoritaire. Ce qui distingue particulièrement cette approche est la méthode de construction séquentielle de cette série : chaque nouveau modèle est ajusté de manière à accorder une importance accrue aux cas que le modèle précédent n'a pas su correctement ajuster ou prédire. De manière intuitive, l'algorithme cible et se concentre sur les cas les plus récalcitrants, tandis que l'intégration des multiples modèles vise à atténuer le risque de surajustement.

Diverses variantes de cet algorithme ont été décrites dans la littérature scientifique, mais la version encapsulée dans la librairie XGBoost (pour eXtreme Gradient Boosting) a remporté un vif succès, particulièrement dans les compétitions de prévisions comme celles organisées sur la plateforme Kaggle.

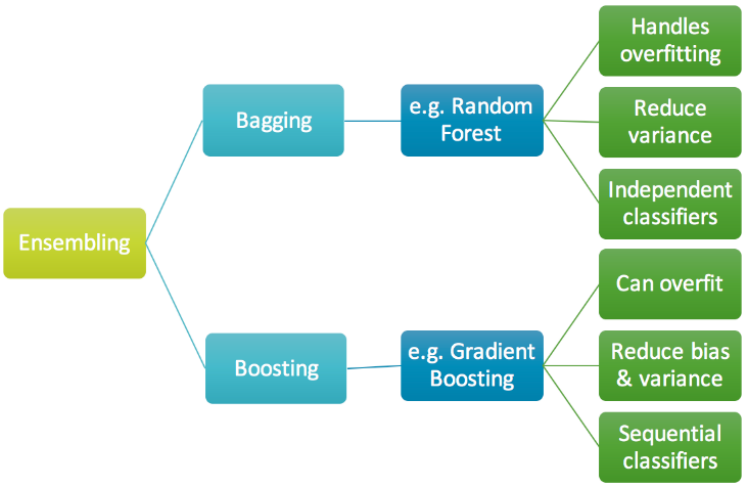


Figure : Eléments de comparaison entre Random Forest et XGBoost

E. Modèle des Réseaux de Neurones

Généralités sur les Réseaux de Neurones

Dans le domaine de l'apprentissage supervisé, les Réseaux de Neurones représentent une méthode de prédiction performante. L'apprentissage supervisé s'applique lorsque l'on dispose de données préalablement classifiées. Ce type d'apprentissage tire parti de l'expérience pour améliorer la capacité de prédiction du modèle. Typiquement, un Réseau de Neurones est structuré en plusieurs couches, où les sorties d'une couche servent d'entrées pour la couche suivante. Ainsi, chaque couche i est constituée de n_i neurones, créant un système hiérarchisé où l'information est progressivement traitée et raffinée.

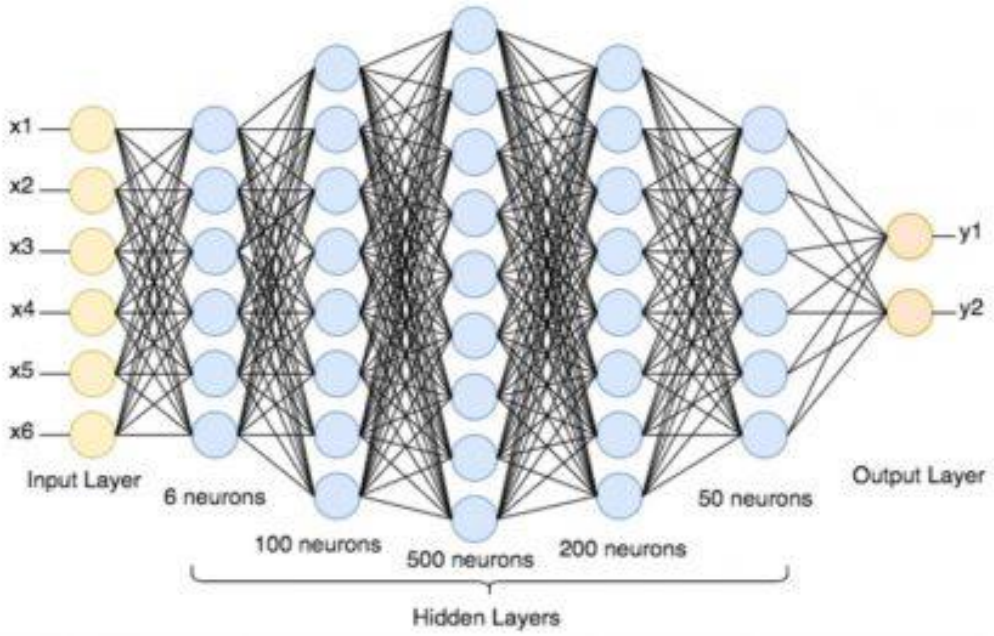


Figure : Schéma d'un Réseau de Neurones

La figure ci-dessus présente un exemple schématique d'un Réseau de Neurones. Nous possédons 6 inputs. Le modèle ici est composé de 5 couches intermédiaires de neurones. La première se composant de 6 neurones et la dernière de 50. Par exemple, à partir d'un historique sur les comportements des assurés (y1 = résiliation et y2 = pas de résiliation), nous entraînons le Réseau de Neurones à partir des observations (paramètres d'entrée). La décision porte sur le comportement que doit réaliser l'assuré : procéder à une résiliation ou laisser son investissement tel quel.

Les Réseaux de Neurones fonctionnent sur le principe de l'induction, apprenant à partir d'expériences spécifiques pour élaborer un mécanisme de décision généralisable. Leur efficacité dépend du volume et de la diversité des situations d'apprentissage, ainsi que de la correspondance entre la complexité des cas traités et celle du problème à solutionner. À l'opposé, bien que les systèmes d'apprentissage symboliques exploitent aussi l'induction, ils le font via une logique algorithmique en enrichissant un corpus de règles déductives.

Un défi majeur des Réseaux de Neurones réside dans leur nature de "boîte noire", rendant leurs processus décisionnels difficiles à interpréter. Les mécanismes internes d'un Réseau de Neurones complexe, conduisant à une décision, restent souvent obscurs pour les concepteurs eux-mêmes. Pour pallier cette opacité, une "neuroscience de l'intelligence artificielle" émerge, visant à démystifier le fonctionnement interne des Réseaux de Neurones. Cette discipline aspire à renforcer la confiance dans les conclusions tirées par ces technologies et les intelligences artificielles qui s'en servent.

Architecture d'un perceptron

L'architecture fondamentale des Réseaux de Neurones repose sur le perceptron, une unité de base pour la modélisation de décisions binaires. Un Réseau de Neurones peut être vu comme un assemblage de

perceptrons travaillant en harmonie pour effectuer des tâches de classification plus complexes ou de régression.

Le perceptron est constitué d'une couche initiale de neurones, chacun représentant une variable d'entrée spécifique, ce qui leur permet d'interpréter les données d'entrée. On y ajoute souvent un neurone de biais, qui est systématiquement actif et transmet une valeur constante de 1, indépendamment des données entrantes. Ces neurones sont connectés à un unique neurone de sortie. Ce dernier reçoit la somme pondérée des signaux des neurones d'entrée, où chaque signal est ajusté par un poids de connexion spécifique.

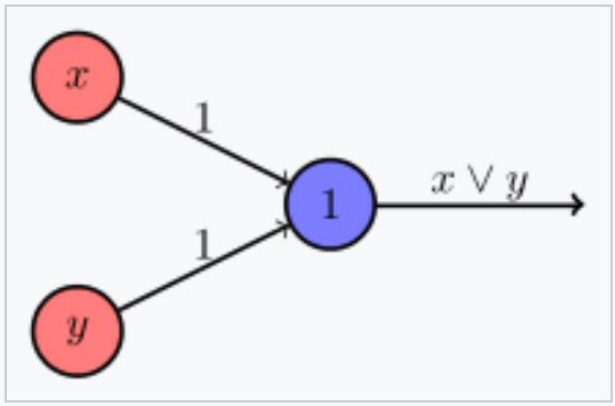


Figure : Illustration d'un perceptron qui calcule le OU logique

Un perceptron simple est le niveau le plus élémentaire d'un Réseau de Neurones. Sa fonction de décision peut être exprimée mathématiquement comme suit :

$$y = f(\sum_{i=1}^{n} w_i x_i + b)$$

où :

- y est la sortie du perceptron,
- x_i représente les entrées,
- w_i sont les poids associés à chaque entrée,
- b est le biais, et
- f est la fonction d'activation, souvent une fonction seuil telle que:

$$f(x) = \begin{cases} 1 \ si \ z \geq 0 \\ 0 \ sinon \end{cases}$$

Cette expression montre comment un perceptron calcule sa sortie en pondérant les entrées, les sommant, y ajoutant un biais, et passant le résultat à travers une fonction d'activation.

Perceptron multicouche (Réseau de Neurones)

Un perceptron multicouche, ou Réseau de Neurones, généralise l'idée du perceptron simple en empilant plusieurs couches de neurones. Chaque couche reçoit en entrée la sortie de la couche précédente, à l'exception de la première couche qui traite les entrées du modèle. Les couches intermédiaires sont appelées couches cachées.

La sortie y d'un Réseau de Neurones multicouche peut être représentée par l'enchaînement des fonctions de décision de chaque couche, noté ici simplement pour une architecture à deux couches cachées :

$$y = f_3(f_2(f_1(\mathbf{x} \cdot \mathbf{W_1} + \mathbf{b_1}) \cdot \mathbf{W_2} + \mathbf{b_2}) \cdot \mathbf{W_3} + \mathbf{b_3})$$

où :

- $\mathbf{x}$ est le vecteur d'entrée,
- $\mathbf{W_n}$ et $\mathbf{b_n}$ représentent respectivement les poids et les biais de la n-ième couche,
- f_n est la fonction d'activation de la n-ième couche, souvent une fonction non linéaire telle que la sigmoïde, la tangente hyperbolique ou la fonction ReLU pour permettre au réseau de modéliser des relations complexes.

La fonction ReLU, pour Rectified Linear Unit, est une fonction d'activation utilisée dans les Réseaux de Neurones, définie par $f(x) = \max(0, x)$. Elle renvoie directement l'entrée si celle-ci est positive et zéro dans le cas contraire. Cette fonction est appréciée pour sa simplicité de calcul et son efficacité dans l'accélération de la convergence du processus d'apprentissage dans les Réseaux de Neurones profonds, en permettant de réduire le problème de disparition du gradient.

Nous examinons par conséquent un modèle paramétrique, car l'expression de la fonction de décision est clairement définie en termes des variables d'entrée. Cependant, il s'agit manifestement d'un modèle non-linéaire.

Dans les architectures complexes des Réseaux de Neurones, il est souvent nécessaire de disposer d'un nombre important d'unités dans les couches cachées, ce qui rend les configurations à une seule couche cachée insuffisantes. Cela souligne l'importance des réseaux neuronaux profonds, qui intègrent plusieurs couches cachées pour améliorer significativement l'efficacité du modèle. Ces structures multicouches permettent d'élaborer des représentations intermédiaires des données, améliorant ainsi la capacité du modèle à apprendre des caractéristiques complexes des données d'entrée.

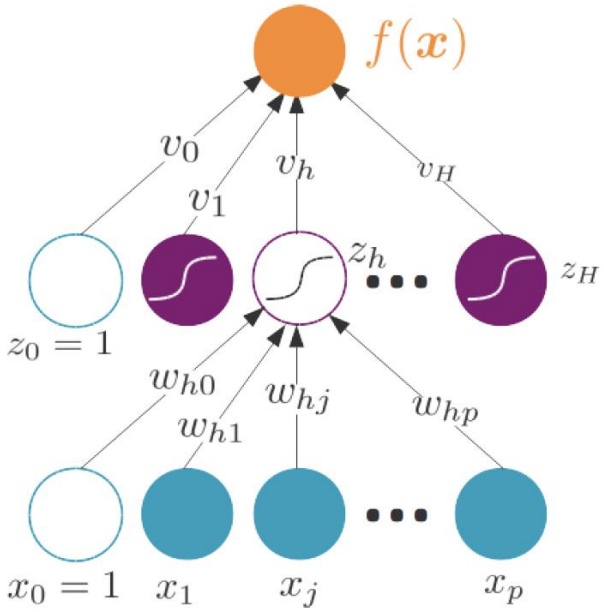


Figure : perceptron multi-couche à une seule couche intermédiaire

La force des Réseaux de Neurones multicouches réside dans leur capacité à générer ces représentations enrichies, facilitant ainsi l'application d'un modèle linéaire (tel que le perceptron) efficace entre l'avant-dernière couche cachée et la couche de sortie.

Concernant la rétropropagation dans les réseaux neuronaux profonds, cette technique s'aligne sur la démarche employée pour le perceptron, s'appuyant sur la méthode de descente de gradient pour optimiser les poids. Le processus alterne entre une phase de propagation avant (Forward Propagation), où les activations des couches intermédiaires sont mises à jour, et une phase de rétropropagation (Backward Propagation), où les gradients des erreurs par rapport aux poids de chaque couche sont calculés en se basant sur les gradients de la couche suivante. La mise à jour des poids intervient durant cette phase descendante, une fois les gradients nécessaires déterminés.

Récapitulatif de la démarche d'entrainement du Réseau de Neurones

La démarche pour entraîner un Réseau de Neurones implique plusieurs étapes clés, visant à optimiser les paramètres du modèle pour réduire l'erreur de prédiction. Voici une analyse détaillée de chaque étape.

i) **Initialisation :** Au commencement, les poids et les biais du réseau sont attribués de manière aléatoire. Cette initialisation aléatoire est cruciale pour briser la symétrie et permettre au réseau d'apprendre diverses fonctions.

ii) **Propagation avant (Forward Propagation) :** Durant cette phase, l'information se propage à travers le réseau depuis l'entrée jusqu'à la sortie. Pour chaque couche, la sortie est calculée en appliquant les poids et les biais à l'entrée, puis en passant le résultat à travers une fonction d'activation. Cette étape permet de déterminer la prédiction du réseau basée sur les entrées actuelles.

iii) **Fonction de Coût :** Après avoir obtenu la sortie du réseau, il est nécessaire d'évaluer la performance du modèle. Une fonction de coût, telle que l'erreur quadratique moyenne pour la régression ou l'entropie croisée pour la classification, mesure l'écart entre les prédictions du modèle et les vraies valeurs cibles. L'objectif est de minimiser cette erreur.

iv) **Rétropropagation (Backward Propagation) :** Cette étape clé permet d'ajuster les poids et les biais en fonction de l'erreur calculée par la fonction de coût. L'erreur est propagée en arrière à travers le réseau, depuis la couche de sortie vers les couches d'entrée. En utilisant le gradient de l'erreur par rapport à chaque paramètre, on ajuste les poids et les biais pour réduire l'erreur.

v) **Itération :** Les étapes de propagation avant et de rétropropagation sont répétées à travers plusieurs itérations, ou époques, jusqu'à ce que le modèle converge vers une solution optimale, c'est-à-dire lorsque l'erreur de prédiction ne s'améliore plus significativement ou atteint un seuil prédéfini d'acceptabilité.

L'algorithme de rétropropagation, qui est le fondement de l'apprentissage des réseaux neuronaux profonds, permet d'ajuster les paramètres des modèles multicouches en prenant bien soin de garder en mémoire qu'on ne sait optimiser explicitement les paramètres du modèle. Par ailleurs, cet outil n'est pas sans limites et requiert une précaution particulière lors de son utilisation.

Problèmes d'optimisation et de minimums locaux : L'un des principaux défis est la stabilité des résultats. De légères variations des paramètres initiaux peuvent entraîner de grandes différences dans les sorties obtenues. Ce phénomène est dû à la topographie complexe de la fonction d'erreur, marquée par des "puits" et "montagnes". Ainsi, le gradient peut se trouver piégé dans un minimum local, aboutissant à des optima d'apparence satisfaisante qui ne sont en réalité que des solutions suboptimales locales. De ce fait, différents choix de valeurs initiales pour les poids peuvent mener à des résultats divergents.

La convergence de l'algorithme et, par extension, la performance de l'apprentissage, dépendent fortement de plusieurs facteurs, notamment le choix des poids initiaux, la normalisation des données (features), la vitesse d'apprentissage (η), et le type de fonction d'activation utilisé. Des stratégies telles que l'application d'apprentissage par mini-batchs ou l'adoption d'une vitesse d'apprentissage adaptative ont envisageables pour atténuer les effets des minimums locaux et favoriser une convergence vers une solution plus globale et robuste.

Problème de Saturation

La saturation d'un modèle se produit lorsque les neurones du réseau, activés par une fonction logistique, affichent majoritairement des sorties proches de 0 ou de 1. Ce phénomène se manifeste lorsque la somme pondérée des signaux d'entrée est excessivement élevée (en valeur absolue), indiquant que les poids de connexion sont trop importants. Dans une telle situation, de légères variations des données d'entrée

n'influencent guère la sortie du réseau, rendant ce dernier incapable d'apprendre davantage. Ce problème est fréquemment causé par des données déséquilibrées, conduisant souvent à un surajustement.

Pour pallier la saturation, il est possible, comme mentionné auparavant, de constituer un ensemble d'entraînement équilibré. Une autre solution consiste à appliquer des techniques de régularisation similaires à celles utilisées en régression. L'introduction d'un terme de pénalisation, de type L1 ou L2, dans la fonction de perte permet de modérer l'ampleur des poids de connexion, contribuant ainsi à atténuer le problème de saturation.

Problème lié au risque d'instabilité du gradient

Le risque lié à l'instabilité du gradient est particulièrement présent dans les réseaux neuronaux à plusieurs couches. Dans ces structures, le gradient tend à diminuer progressivement lorsqu'on remonte des neurones de sortie vers ceux d'entrée, ce qui entraîne un apprentissage plus lent pour les neurones des premières couches par rapport à ceux situés plus proches de la sortie. Ce phénomène est connu sous le nom de "vanishing gradient", ou disparition du gradient. À l'inverse, il est également possible que le gradient augmente de manière excessive dans les couches basses, ce qui est décrit comme le problème de "l'explosion du gradient" ou "exploding gradient".

3. Sélection du modèle de prédiction

A. Initialisation des modèles et entrainement

Dans le cadre de notre étude visant à prédire les résiliations de polices d'assurance vie en UC en fonction de l'évolution des caractéristiques de la clientèle et des indicateurs économiques et financier, nous initialisons plusieurs modèles. Selon le cas, certains modèles comme le Random Forest ou les Réseaux de Neurones peut s'avérer plus performant selon les cas. Pour information, les modèles de prédiction sur régressions linéaires pénalisées Ridge et Lasso ont été abandonnés puisque nous traitons un problème de classification. Chaque modèle est entrainé sur des données d'entraînement normalisées pour assurer une évaluation équitable et optimiser leur capacité prédictive.

Après l'entrainement des modèles, chacun d'entre eux sont testés sur un ensemble de données de test pour évaluer leur performance. Nous calculons les métriques clés telles que l'exactitude (accuracy), la précision (precision), le rappel (recall), le score F1 et l'aire sous la courbe ROC (AUC) pour chaque modèle. Les matrices de confusion pour chaque modèle seront calculées séparément. Ils nous fourniront une vue détaillée de la performance des modèles en termes de vrais positifs, faux positifs, vrais négatifs et faux négatifs. Ces métriques seront une base solide pour comparer l'efficacité des différents modèles et sélectionner par la suite le modèle le plus pertinent pour notre prédiction.

a. Performance des modèles

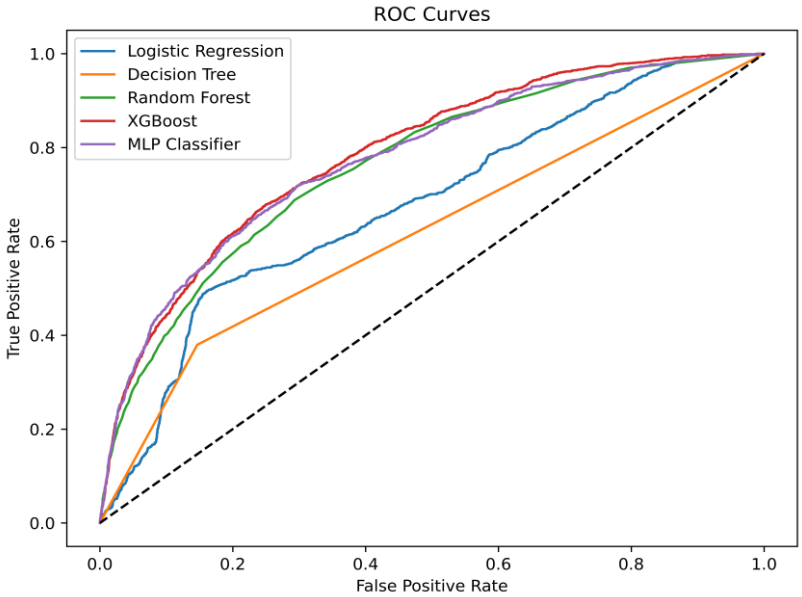


Figure : Courbes ROC des modèles entrainés et testésS

Les courbes ROC offrent une représentation graphique du compromis entre le taux de vrais positifs et le taux de faux positifs pour différents modèles prédictifs. Dans cette représentation, le modèle XGBoost se distingue particulièrement, sa courbe s'approchant davantage du coin supérieur gauche, ce qui signifie une performance supérieure en pour ce modèle. Cela souligne la capacité supérieure du modèle XGBoost à différencier les classes prédites. Cette performance indique que le modèle est particulièrement efficace pour distinguer les cas de résiliation des polices d'assurance de ceux qui ne le sont pas. Parallèlement, les modèles Random Forest et Réseaux de Neurones MLP (Multilayer Perceptron) ont également affiché des résultats prometteurs, avec des scores AUC élevés, témoignant de leur robustesse.

En revanche, les modèles de Régression Logistique et d'Arbre de Décision ont montré une efficacité moindre en comparaison. Bien qu'ils fournissent des bases solides pour la compréhension des dynamiques sous-jacentes, leur performance en termes de capacité à distinguer précisément entre les différentes classes est inférieure à celle observée pour les modèles plus complexes.

Aussi, l'analyse des courbes ROC confirme que la régression logistique est exploitable avec des indices économiques. Toutefois, plutôt que d'utiliser directement des variables brutes comme les taux d'intérêt, il est préférable de construire des indicateurs plus informatifs. Par exemple, le spread de taux (écart entre le taux actuel et le taux de souscription) permettrait de mieux capter l'effet des variations économiques sur les résiliations.

Pour une sélection avisée des modèles, il est primordial de visualiser les autres scoring de performances :

Modèle	Accuracy	Precision	Recall	F1-score	AUC

Régression Logistique	0.6244	0.25	0.002	0.0039	0.6787
Arbres de Décision	0.7706	0.3545	**0.3797**	0.3667	0.6166
Forêts Aléatoires	0.8281	0.5136	0.3209	**0.395**	0.7643
XGBoost	**0.8426**	**0.6073**	0.283	**0.3861**	**0.7873**
Réseaux de Neurones	**0.8414**	**0.6076**	0.2621	0.3662	**0.7772**

NB : Il serait intéressant de mentionner que, dans l'optique d'une économie de temps au regard du contexte transactionnel, l'utilisation d'un AutoML (Automated Machine Learning) pourrait permettre d'automatiser l'évaluation et l'optimisation des modèles afin d'identifier rapidement celui offrant les meilleures performances sur le dataset.

Logistic Regression

L'analyse des performances du modèle de Régression Logistique révèle plusieurs points critiques concernant sa capacité prédictive :

- **Accuracy (Précision globale) :** La précision globale du modèle de Régression Logistique est considérée comme relativement bonne. Cependant l'analyse des autres scores montre que ce modèle n'est pas forcément le meilleur. En effet, le dataset contenant une grande majorité de cas négatifs (par exemple, des contrats qui ne sont pas résiliés) et une petite minorité de cas positifs (contrats résiliés), le modèle peut atteindre une haute précision globale simplement en prédisant la classe majoritaire pour la majorité des observations. Cela signifie que même un modèle peu sophistiqué ou peu performant peut afficher une haute précision globale si les classes sont fortement déséquilibrées.
- **Precision (Précision des prédictions positives) :** La précision très faible signale que, parmi les instances classées comme positives par le modèle, seulement un petit nombre sont réellement positives. Dans le contexte de la prédiction de résiliation, cela signifie que le modèle a tendance à surévaluer le risque de résiliation, ce qui pourrait conduire à des interventions inutiles ou à des évaluations erronées du risque (risque de sous-évaluation de la valeur du portefeuille).
- **Recall (Rappel ou sensibilité) :** Un rappel extrêmement faible indique que le modèle est inefficace pour détecter les vrais cas de résiliation. Cela est particulièrement problématique car cela signifie que de nombreux vrais positifs, c'est-à-dire des résiliations effectives, ne sont pas identifiés. Dans une application transactionnelle, cela pourrait entraîner une sous-estimation du risque de résiliation, soit une surévaluation de la valeur du portefeuille.
- **F1-score :** Le score F1 est critique car il combine la précision et le rappel en une seule métrique. Un score F1 très faible reflète les performances insuffisantes du modèle tant en termes de précision que de rappel, soulignant son incapacité à fournir des prédictions équilibrées et fiables.

- **AUC (Area Under the ROC Curve) :** Un score AUC modéré indique une capacité limitée du modèle à différencier les classes de résiliation et de non-résiliation. Bien que supérieure à une performance aléatoire, cette valeur suggère que le modèle possède une utilité limitée pour prédire avec précision les risques de résiliation, surtout lorsqu'il est crucial de distinguer clairement entre les différents résultats possibles. Ce score est en phase avec la courbe ROC.

Ainsi, bien que la Régression Logistique soit un outil populaire et généralement efficace pour les problèmes de classification binaire, son application dans ce contexte spécifique montre des limites significatives.

Cependant ces performances proviennent probablement d'un feature engineering insuffisant. En effet, au-delà de l'encodage de variables catégorielles pour les transformer en données qualitatives, un travail plus approfondi de feature engineering aurait pu être appliqué aux variables économiques, plutôt que de les utiliser sous leur forme brute. En particulier, la création d'indices économiques plus pertinents, tels que le spread de taux (écart entre le taux à la date d'analyse et le taux à la souscription), aurait permis de mieux capter l'impact des conditions économiques sur la résiliation des contrats. Ce type de transformation aurait pu améliorer la capacité prédictive des modèles en reflétant plus fidèlement les effets des évolutions économiques.

Desicion Tree (Arbre de décision)

L'analyse des performances du modèle d'arbre de décision révèle des améliorations notables par rapport à la Régression Logistique, bien que certaines limites demeurent :

- **Accuracy (Précision globale) :** La précision du modèle d'arbre de décision est considérée bonne dans l'ensemble. Toutefois, cette précision reste modeste, indiquant que des erreurs de classification subsistent et que le modèle pourrait ne pas être suffisamment fiable pour toutes les applications décisionnelles.
- **Precision (Précision des prédictions positives)** : L'augmentation significative de la précision indique que le modèle d'arbre de décision est plus apte à identifier correctement les cas de résiliation réels parmi les prédictions positives. Cette amélioration est cruciale dans le contexte de la prédiction de résiliation, car elle réduit le risque d'actions inappropriées basées sur des prédictions erronées.
- **Recall (Rappel ou sensibilité) :** L'amélioration du rappel signifie que le modèle d'arbre de décision est plus compétent pour détecter les vrais positifs parmi tous les cas réels de résiliation. Cela est particulièrement important pour minimiser le risque de manquer des résiliations effectives, ce qui pourrait avoir des implications significatives dans notre contexte.
- **F1-score :** Un score F1 plus élevé reflète une meilleure performance globale du modèle, indiquant un équilibre amélioré entre la précision et le rappel. Cette métrique est essentielle pour

évaluer l'efficacité globale du modèle dans un contexte où il est important de maintenir un bon équilibre entre identifier correctement les résiliations et éviter les fausses alertes.

- **AUC (Area Under the ROC Curve) :** Bien que l'AUC pour l'arbre de décision soit légèrement inférieure à celle obtenue avec la Régression Logistique, elle indique toujours une capacité raisonnable à différencier entre les résiliés et les non-résiliés. Cette légère baisse peut refléter des compromis inhérents à la méthode de l'arbre de décision, notamment en termes de généralisation.

En résumé, le modèle d'arbre de décision montre des améliorations significatives dans plusieurs métriques clés par rapport à la Régression Logistique, suggérant qu'il pourrait être plus adapté pour certaines applications dans la prédiction de résiliation. Néanmoins, la légère baisse de l'AUC souligne l'importance de continuer à évaluer et comparer différents modèles pour identifier celui qui offre la meilleure performance selon les critères spécifiques du contexte d'application.

Random Forest (Forêts Aléatoires)

L'analyse des performances du modèle Random Forest révèle une amélioration notable par rapport aux modèles précédents, notamment en termes de précision et de capacité discriminatoire :

- **Accuracy (Précision globale) :** la meilleure précision globale avec le modèle Random Forest indique que ce modèle est capable de classer correctement un pourcentage plus élevé d'observations, tant positives que négatives, par rapport aux modèles antérieurs. Cette amélioration reflète la capacité du Random Forest à intégrer les informations de multiples arbres de décision, réduisant ainsi le risque de surajustement et augmentant la robustesse du modèle.
- **Precision (Précision des prédictions positives)** : Une précision significativement améliorée signale que, parmi les instances que le modèle Random Forest a classées comme résiliations, une plus grande proportion est effectivement correcte. Cela indique que le modèle est plus fiable et plus précis dans l'identification des contrats susceptibles d'être résiliés, ce qui est crucial dans notre cas.
- **Recall (Rappel ou sensibilité) :** Un rappel légèrement inférieur par rapport au modèle d'arbre de décision peut indiquer que le Random Forest, malgré sa précision accrue, pourrait manquer certains vrais positifs. Cependant, cette légère baisse est souvent un compromis acceptable pour une augmentation significative de la précision et de la spécificité globale. Néanmoins, dans notre contexte transactionnel, nous cherchons avant tout à nous couvrir contre un risque de sous-évaluation des résiliations (ce qui revient à surévaluer la valeur du portefeuille), nous chercherons donc à minimiser le plus possible le fait de manquer des vrais positifs.
- **F1-score (Score F1) :** L'amélioration du score F1 avec le modèle Random Forest démontre un meilleur équilibre entre la précision et le rappel, ce qui est essentiel pour les modèles où il est

important de maintenir un bon niveau de performance sur les deux fronts, particulièrement dans des contextes où les coûts des faux positifs et des faux négatifs sont tous deux significatifs.

- **AUC (Area Under the ROC Curve) :** Une AUC considérablement plus élevée pour le Random Forest indique une excellente capacité du modèle à différencier les classes de résiliation et de non-résiliation. Une valeur élevée d'AUC est particulièrement importante dans les contextes où les distinctions fines entre les probabilités de classe sont cruciales pour la prise de décision.

En conclusion, le modèle Random Forest se distingue par sa robustesse et sa capacité à fournir des prédictions fiables et nuancées, ce qui le rend particulièrement adapté pour l'analyse du risque de résiliation. Sa performance supérieure en termes de précision, de score F1 et d'AUC en fait un choix privilégié pour les analyses où la qualité et la fiabilité des prédictions sont primordiales.

XGBoost :

L'analyse des performances du modèle XGBoost met en évidence son efficacité exceptionnelle par rapport aux autres modèles testés :

- **Accuracy (Précision globale) :** le modèle XGBoost affiche la meilleure précision globale parmi tous les modèles évalués, ce qui indique une capacité supérieure à classer correctement une grande majorité des observations. Cette haute précision témoigne de l'efficacité de l'algorithme du modèle XGBoost, qui combine plusieurs arbres de décision de manière optimisée pour réduire l'erreur et éviter le surajustement.
- **Precision (Précision des prédictions positives) :** Avec la précision la plus élevée, le modèle XGBoost démontre sa fiabilité dans l'identification des résiliations. Cela signifie que, parmi les cas prédits comme étant des résiliations, la majorité sont correctement identifiés, ce qui est crucial pour les applications nécessitant une grande confiance dans les prédictions positives, comme les stratégies de rétention ciblées ou la fiabilité de valorisation du portefeuille
- **Recall (Rappel ou sensibilité) :** Bien que le rappel soit légèrement inférieur à celui observé pour le modèle Random Forest, il reste élevé pour le modèle XGBoost, indiquant que le modèle est capable d'identifier un grand nombre de vrais positifs. Cela est essentiel pour minimiser le risque de manquer des cas de résiliation qui pourraient avoir des implications financières significatives.
- **F1-score (Score F1) :** Le score F1 le plus élevé pour le modèle XGBoost reflète une performance équilibrée entre la précision et le rappel, soulignant l'efficacité du modèle à maintenir un bon niveau de performance sur ces deux aspects critiques. Un score F1 élevé est indicatif d'un modèle bien ajusté qui équilibre correctement le compromis entre identifier correctement les résiliations et minimiser les fausses alertes.
- **AUC (Area Under the ROC Curve)** : Avec l'AUC la plus élevée, le modèle XGBoost démontre une capacité indéniable à distinguer entre les classes positives et négatives, ce qui est

un atout majeur pour les décisions basées sur la probabilité de résiliation. Une AUC élevée indique que le modèle est capable de classer correctement les observations avec une grande confiance sur distinguant efficacement les probabilités de résiliation des non-résiliations sur un large éventail de seuils. Cela illustre la compétence du modèle à opérer dans des contextes où la distinction fine entre les classes est essentielle pour la prise de décision.

En résumé, XGBoost se distingue comme le modèle le plus performant dans l'ensemble des métriques évaluées, offrant une solution robuste et fiable pour la prédiction des résiliations dans les portefeuilles d'assurance vie en UC. Sa capacité à fournir des prédictions précises et fiables en fait un outil précieux pour les analyses prédictives et la prise de décision stratégique dans le contexte des assurances.

MLP neural networks (Réseaux de Neurones)

L'évaluation des performances du MLP (Multilayer Perceptron) montre que ce modèle de Réseau de Neurones offre des résultats très compétitifs, comparables à ceux du modèle XGBoost :

- **Accuracy (Précision globale) :** Le MLP présente une précision globale très élevée, similaire à celle observée pour le modèle XGBoost, ce qui indique que le modèle est capable de classer correctement un pourcentage élevé des observations, tant positives que négatives. Cette haute précision suggère que le MLP est efficace dans l'ensemble des situations, fournissant des prédictions fiables et robustes.
- **Precision (Précision des prédictions positives) :** Bien que la précision du MLP soit légèrement inférieure à celle du modèle XGBoost, elle demeure élevée, indiquant que le modèle est fiable dans l'identification des résiliations. Cela signifie que, parmi les cas prédits comme résiliations, une proportion importante est correctement identifiée.
- **Recall (Rappel ou sensibilité) :** Le rappel du MLP est comparable à celui du modèle XGBoost, indiquant que le modèle est capable d'identifier un grand nombre de vrais positifs. Cette capacité à détecter les résiliations réelles est cruciale pour éviter de passer à côté de cas importants qui pourraient avoir des conséquences financières.
- **F1-score (Score F1) :** Le score F1 du MLP, bien qu'un peu inférieur à celui du modèle XGBoost, reste très élevé, témoignant d'une bonne balance entre la précision et le rappel. Ce score élevé indique que le MLP maintient un équilibre efficace entre la détection des résiliations et la limitation des fausses alertes.
- **AUC (Area Under the ROC Curve) :** L'AUC du MLP, légèrement inférieure à celle de du modèle XGBoost, demeure très élevée, démontrant une excellente capacité du modèle à différencier les classes de résiliation et de non-résiliation. Une AUC élevée est indicative d'une performance solide sur l'ensemble des seuils de décision, ce qui est particulièrement utile dans les contextes où les décisions doivent être prises à différents niveaux de probabilité.

En conclusion, le modèle de Réseau de Neurones MLP se révèle être un outil puissant et efficace pour la prédiction des résiliations, offrant des performances globales très élevées et comparables à celles du modèle XGBoost. Sa capacité à fournir des prédictions précises et son équilibre entre les différentes métriques de performance en font une option viable pour les analyses prédictives dans notre contexte.

Analyse des matrices de confusions

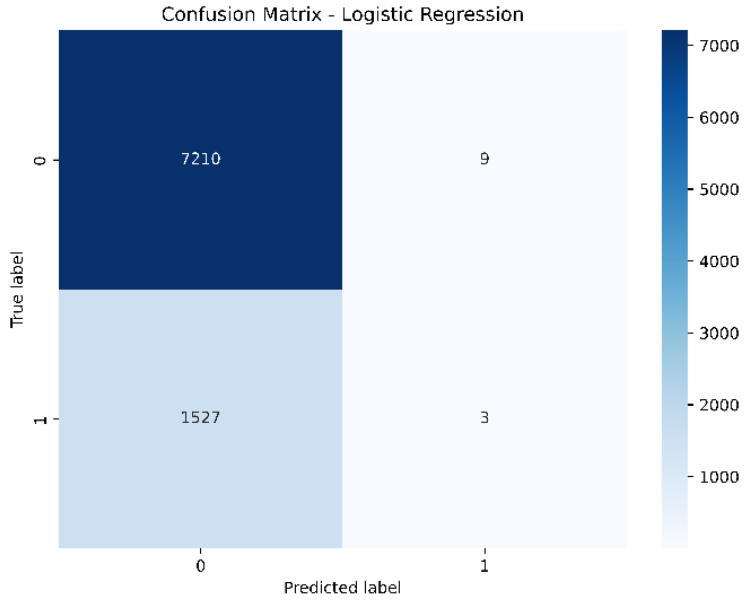


Figure : Matrices de confusion pour Logistic Regression

Logistic Regression : la matrice montre un nombre élevé de vrais négatifs (TN), mais très peu de vrais positifs (TP), indiquant que le modèle a tendance à prédire la classe négative. Cela se traduit par un faible rappel (le modèle ne capte qu'une petite fraction des résiliations effectives) et une faible précision (tendance à générer de nombreux faux positifs diluant ainsi la fiabilité des prédictions de résiliation) pour la classe positive. Ces résultats sont cohérents avec les scores de performance précédemment discutés.

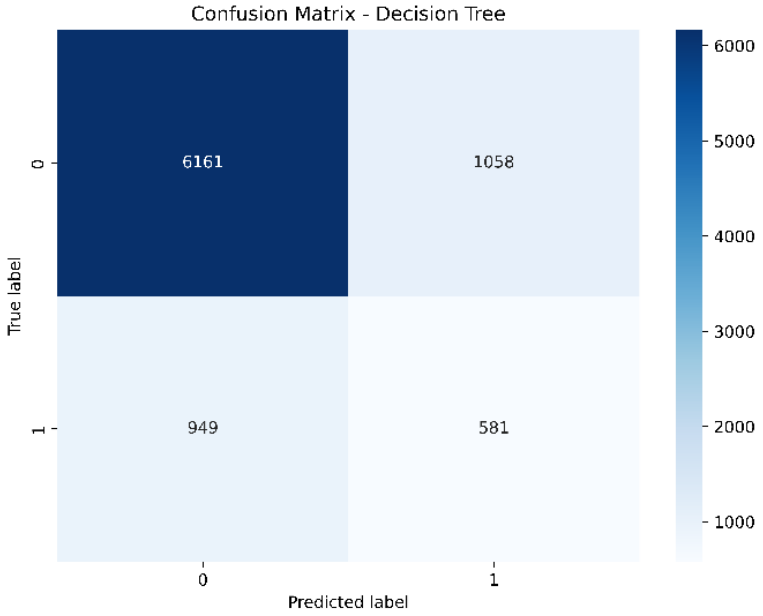


Figure : Matrices de confusion pour Decision Tree

Decision Tree : cette matrice présente un meilleur équilibre entre les TP et TN par rapport à la Régression Logistique, ce qui suggère que le modèle bénéficie d'une meilleure sensibilité et spécificité globales. Cette amélioration peut être attribuée à la nature même des arbres de décision, qui segmentent l'espace des caractéristiques de manière plus granulaire. Cependant, il y a un nombre significatif de faux positifs (FP) et de faux négatifs (FN), ce qui indique une certaine confusion dans les prédictions du modèle. En effet, le modèle d'arbre de décision serait souvent confus dans ses classifications, attribuant incorrectement des étiquettes positives à des négatifs et vice versa. Cette confusion peut résulter de la surcomplexité de l'arbre ou de son incapacité à généraliser à partir des données d'entraînement. Cela se reflète dans une précision et un rappel modérés.

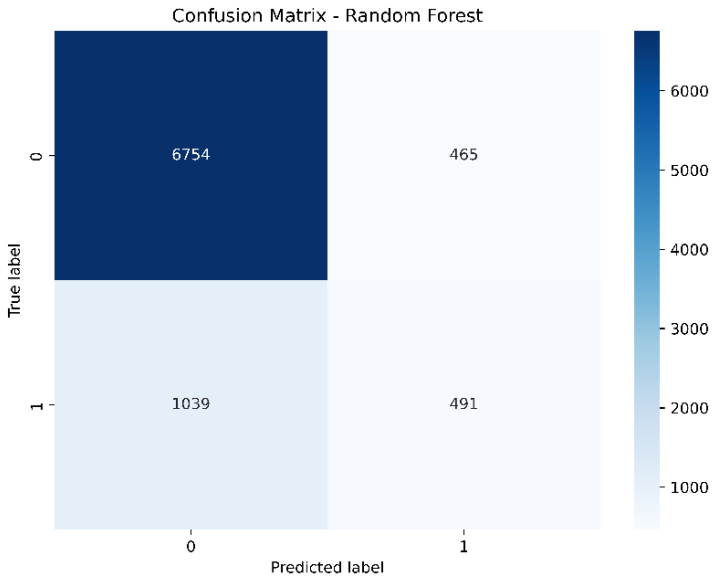


Figure : Matrices de confusion pour Random Forest

Random Forest : le modèle Random Forest montre une amélioration notable en termes de TP et TN au global, avec moins de FP et FN par rapport au modèle Decision Tree. Cela indique une meilleure capacité à classer correctement les instances notamment grâce à son approche d'agrégation d'arbres de décision (bagging), se traduisant par une meilleure précision, un meilleur rappel, et donc un meilleur score F1. L'amélioration de la précision du modèle indique qu'il est plus fiable lorsqu'il prédit une résiliation. L'amélioration du rappel témoigne que le modèle est capable de détecter une proportion plus élevée des résiliations réelles parmi toutes les résiliations potentielles.

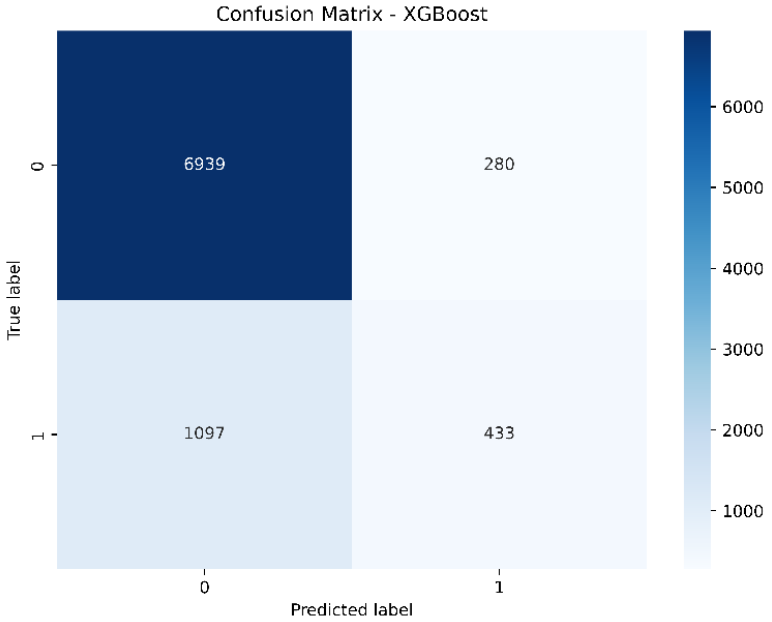


Figure : Matrices de confusion pour XGBoostma

XGBoost : la matrice de confusion pour le modèle XGBoost révèle le meilleur équilibre entre TP, TN, FP, et FN parmi tous les modèles. Cela indique une bonne capacité à distinguer entre les classes, ce qui est corroboré par de meilleurs scores de performance (précision, rappel, score F1) et un AUC élevé. Le modèle XGBoost maintient un bon équilibre entre la détection des résiliations et la limitation des classifications incorrectes. Par ailleurs, un rééchantillonnage des données (oversampling) ou une pondération plus importante des résiliations pour compenser le déséquilibre aurait pu aider à obtenir une meilleure détection des résiliations tout en limitant les erreurs de classification.

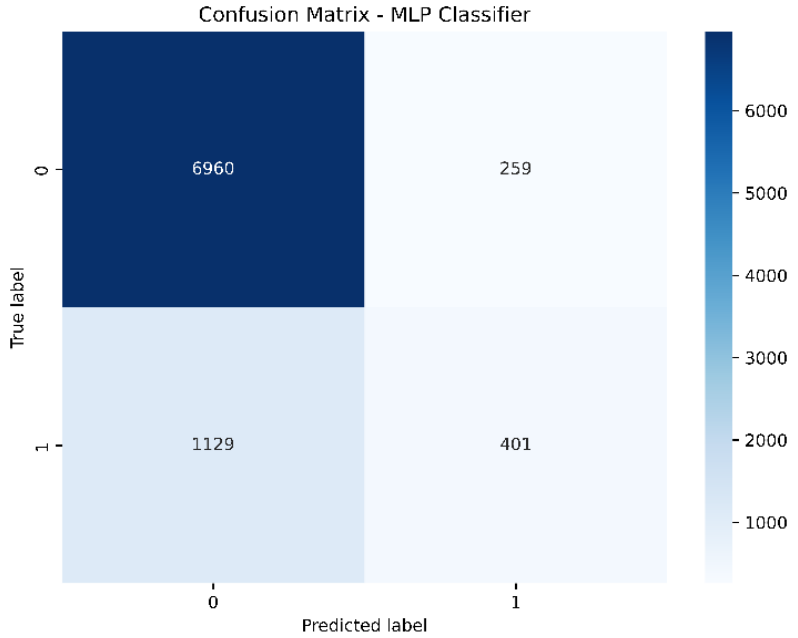


Figure : Matrices de confusion pour le Réseau de Neurones MLP

Réseau de Neurones MLP : la matrice pour le MLP Classifier montre également de bonnes performances, avec un nombre élevé de TP et TN. Cependant, il y a toujours une présence de FP et FN, bien que moins prononcée que dans les modèles de Decision Tree et Random Forest, indiquant une bonne capacité de classification. Par rapport au modèle XGBoost, le modèle ne surdiagnostique pas les résiliations (moins de FP), mais manque moins de cas réels de résiliation (moins de FN), donc moins intéressant dans notre contexte transactionnel où nous souhaitons manquer le moins de cas réels de résiliation pour limiter le risque de survaleur du portefeuille.

B. Sélection du modèle de prédiction

Au regard de l'ensemble des critères de performance analysés (F1 score, précision, AUC score, rappel, courbe ROC, matrices de confusion), le modèle XGBoost se distingue comme le plus fiable pour la prévision du risque de résiliation.

Voici les raisons principales de cette préconisation :

- **F1 Score :** Le modèle XGBoost a montré un F1 score élevé, indiquant un excellent équilibre entre la précision et le rappel. Cela signifie que le modèle est capable de prédire correctement un grand nombre de cas positifs tout en limitant le nombre de faux positifs.
- **Précision et Rappel :** Le modèle XGBoost a également démontré une précision et un rappel élevés, ce qui est crucial pour les applications où il est important de minimiser les erreurs de classification, comme dans la prévision du risque de résiliation.

- **AUC Score :** Avec un AUC score très élevé, le modèle XGBoost a prouvé sa capacité à distinguer efficacement entre les classes positives et négatives. Un AUC élevé est un indicateur de la robustesse du modèle face à différentes seuils de classification.
- **Courbe ROC :** La courbe ROC du modèle XGBoost, se rapprochant plus de l'angle supérieur gauche, illustre sa supériorité en termes de taux de vrais positifs par rapport au taux de faux positifs. Cela témoigne de sa capacité à maximiser les vrais positifs tout en minimisant les faux positifs.
- **Matrices de Confusion :** Les matrices de confusion ont révélé que le modèle XGBoost a le meilleur équilibre entre les vrais positifs (TP) et les vrais négatifs (TN), avec un nombre réduit de faux positifs (FP) et de faux négatifs (FN). Cela indique une excellente capacité de classification correcte des instances.

Le modèle XGBoost se distingue donc par sa performance globale supérieure sur tous les critères évalués. Sa capacité à fournir des prédictions précises et fiables, avec un meilleur équilibre entre les différents indicateurs de performance, en fait le choix le plus recommandé pour la prévision du risque de résiliation.

C. Les caractéristiques importantes dans la prédiction du modèle (feature importances)

Les caractéristiques importantes (feature importances) dans les modèles de prédiction désignent le degré d'influence que chaque variable explicative a sur la prédiction du modèle. En d'autres termes, elles indiquent à quel point chaque caractéristique contribue à la capacité du modèle à faire des prédictions précises. Cette information est cruciale pour comprendre le comportement du modèle, pour identifier les variables qui jouent un rôle majeur dans la prédiction et, par conséquent cela permet d'interpréter les résultats du modèle de manière éclairée. En comprenant l'importance de ces variables, les assureurs et les analystes peuvent mieux appréhender les facteurs clés qui influencent les décisions de résiliation.

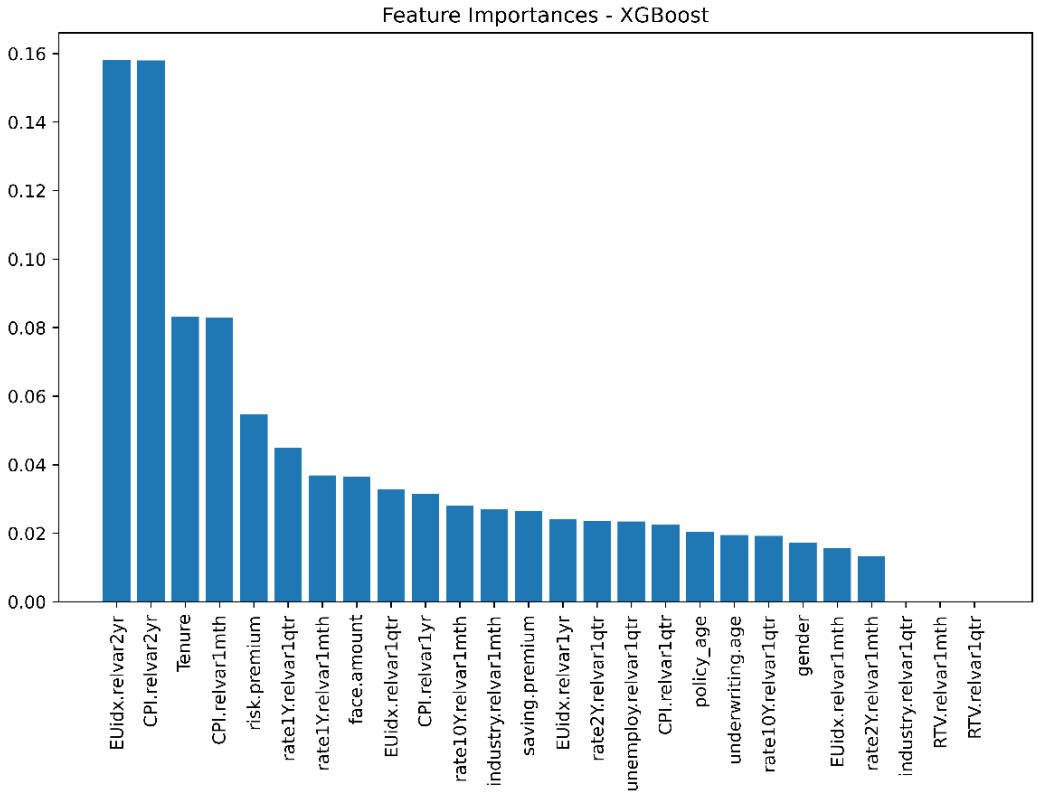


Figure : Caractéristiques importantes (*Feature importances*) du modèle XGBoost entrainé

Dans le contexte de la prédiction du risque de résiliation pour un portefeuille d'assurance vie en UC avec le modèle XGBoost, les variables suivantes ont été identifiées comme les plus influentes :

- **EUidx.relvar2yr** (Variation relative de l'indice boursier européen sur deux ans) : cette variable reflète les tendances économiques à long terme qui peuvent influencer les décisions de résiliation des polices. Une variation significative peut indiquer des changements dans le climat économique qui affectent la stabilité financière ou la perception du risque par les assurés.
- **CPI.relvar2yr** (Variation relative de l'Indice des Prix à la Consommation sur deux ans) : Cette mesure indique comment l'inflation, ou la variation des coûts de la vie, a évolué sur deux ans. Les changements dans l'IPC peuvent influencer le pouvoir d'achat des consommateurs et, par conséquent, leur capacité ou leur volonté de continuer à payer les primes d'assurance, impactant ainsi le risque de résiliation.
- **Tenure (Ancienneté de la relation avec la police) :** L'ancienneté de la relation avec la police est un indicateur évident du risque de résiliation. Plus il est élevé, plus le risque de résiliation est élevé (mesure empirique).
- **CPI.relvar1mth (Variation relative de l'Indice des Prix à la Consommation sur un mois) :** Cette variable capture les fluctuations à court terme de l'inflation, offrant un aperçu des conditions économiques immédiates. Des variations brusques peuvent signaler des événements

économiques qui pourraient pousser les assurés à reconsidérer leurs engagements financiers, y compris les assurances vie.

- **Risk.premium (Prime de risque de souscription) :** La prime de risque représente le montant additionnel payé pour assumer un risque plus élevé. Dans le contexte de l'assurance, elle reflète le niveau de risque associé à un contrat spécifique. Une prime de risque élevée peut être un indicateur exposition des investissements de l'assuré aux risques de marché.

Dans le cadre de notre analyse qui vise à prédire les résiliations de contrats d'assurance vie UC, l'attention sera donc particulièrement portée sur ces variables identifiées comme influentes par le modèle XGBoost.

L'objectif est de capitaliser sur cette compréhension pour anticiper les résiliations potentielles dans le futur.

IV. Application des prédictions et écoulement du portefeuille

La démarche globale que nous allons adopter vise à prédire les résiliations futures au sein d'un portefeuille d'assurance vie en UC pour estimer l'écoulement du portefeuille. L'écoulement du portefeuille fait référence à la diminution progressive du nombre de contrats actifs dans le portefeuille, sans considérer l'ajout de nouveaux contrats (new business). Cette approche permet d'évaluer la valeur résiduelle du portefeuille existant en se concentrant uniquement sur les contrats en cours à la date d'étude. Ici nous prendrons aussi pour hypothèse qu'il n'y a pas versements libres ou programmés dans les projections futures pour ne pas complexifier notre étude.

Base de préparation sur le portefeuille à valoriser dans le cadre de la transaction :

Pour notre analyse, nous utilisons les données historiques couvrant les entrées et sorties de portefeuille de 1999 à 2007. Les contrats qui sont encore actifs à la fin de 2007 constituent le portefeuille à valoriser. Cette base de données nous permettra de comprendre les tendances de résiliation passées. L'idée sera d'extrapoler les tendances historiques pour estimer l'écoulement futur du portefeuille, c'est-à-dire l'érosion du portefeuille liés aux résiliations, sans l'intégration de nouveaux contrats.

Projection des variables économiques et financières, et des variables intrinsèques des assurés :

Nous allons projeter les variables économiques et financières telles que le CPI et l'EUidx pour anticiper leur influence sur le comportement de résiliation des assurés. Les variations de ces indices sur deux ans et un mois (CPI.relvar2yr, CPI.relvar1mth, et EUidx.relvar2yr) ainsi que les caractéristiques intrinsèques des contrats (tenure et risk.premium) sont des indicateurs clés qui influenceront nos prévisions de résiliation.

Application du modèle pour prédire les résiliations :

Le modèle XGBoost, entraîné sur les données historiques, sera appliqué pour prédire les taux de résiliation sur les périodes futures. Ce modèle, entrainé sur la base des comportements clients observés des données historiques, permettra de fournir une trajectoire de résiliation selon l'évolution des différents indicateurs de marché et caractéristiques du portefeuille, et donc d'estimer les résiliations attendues dans le portefeuille.

Estimation de la valeur des flux futurs de marge (PVFP – Present Value of Future Profit) puis de la valeur économique du portefeuille :

Après avoir estimé l'écoulement du portefeuille, il conviendrait de calculer la valeur économique du portefeuille. Nous procèderons à une approche simplifiée qui consiste à appliquer les taux de marge nette par contrat aux contrats restants dans le portefeuille projeté. Cette étape finale permettra de quantifier la valeur future des marges générées par le portefeuille, offrant ainsi une estimation de sa valeur intrinsèque.

En résumé, cette démarche méthodique nous permettra de fournir une estimation éclairée de l'évolution future du portefeuille et de sa valeur, en se basant sur une analyse pertinente des tendances de résiliation et des facteurs économiques et intrinsèques influençant le comportement des assurés.

L'atout majeur de cette estimation réside dans sa capacité à refléter fidèlement le comportement des assurés du portefeuille concerné par la transaction.

Habituellement, dans un cadre transactionnel, l'acquéreur n'a pas accès aux lois de résiliation de l'entité cible. Toutefois, il est courant d'obtenir une datatape qui compile l'historique des données du portefeuille, offrant ainsi une base solide pour l'analyse et la prédiction par le biais des méthodes de machine learning.

1. Projection des données d'input du modèle

A. Le portefeuille étudié

Nous préparons la base de données pour extraire uniquement les données relatives au portefeuille à la date de valorisation, à savoir contrats qui sont encore actifs à la fin de 2007.

Voici les statistiques principales sur la clientèle issue du dataset (portefeuille faisant l'objet de la transaction) :

- Nombre de polices : 10,751
- Âge moyen de souscription : 51.02 ans
- Âge moyen de la police : 53.14 ans
- Proportion d'hommes : 65%
- Proportion de femmes : 35%
- Tenure moyenne : 2.19 ans

Comme précisé au début, il est important de prendre en compte le fait que ce portefeuille étudié est issu de 8 générations de collectes et donc n'est pas totalement représentatif d'un portefeuille mature. Ici les

collectes sur les dernières années n'ont pas encore été « écoulées », ce qui tire l'ancienneté moyenne à la baisse (2.2 ans d'ancienneté moyenne).

B. Projection des caractéristiques importantes (feature importances)

Dans le processus de valorisation d'un portefeuille d'assurance, il est courant de réaliser des projections sur une longue période, souvent entre 50 et 70 ans, pour capturer l'ensemble des flux futurs et évaluer la valeur actuelle nette de ces flux. Cependant, par souci de simplification de notre analyse, **nous limiterons nos projections à 15 ans**.

Les variations des indices économiques et financiers tels que EUidx.relvar2yr (variation relative d'un indice boursier européen sur deux ans), CPI.relvar2yr et CPI.relvar1mth (variation relative du Consumer Price Index sur deux ans et sur un mois), ainsi que les caractéristiques propres aux contrats d'assurance comme l'ancienneté de la relation (Tenure) et la prime de risque (risk premium), sont des facteurs déterminants pour anticiper les comportements de résiliation des assurés.

Pour les données intrinsèques des assurés, telles que l'âge et l'ancienneté (Tenure), nous adopterons une approche incrémentale simple, étant donné que ces attributs évoluent de manière prévisible année après année. De même, nous choisirons de conserver constant le niveau de risque associé à chaque assuré pour ne pas complexifier notre étude.

En ce qui concerne les autres variables explicatives telles que les variables économiques et financières dans notre cas, il est courant de réaliser des projections stochastiques sur l'horizon de projection. Ces scénarios stochastiques nous permettent d'évaluer les comportements de résiliation sous différents scénarios économiques et d'estimer ainsi un écoulement moyen du portefeuille.

Les assureurs disposent en général de **Générateurs de Scénarios Economiques (GSE)** permettant de fournir les scénarios d'actifs nécessaires aux calculs du Best Estimate et des charges en capital règlementaires, mais aussi de remplacer les études ALM jusqu'alors déterministes par des modèles stochastiques reflétant les évolutions historiques des indices et taux considérés.

Ainsi dans le contexte transactionnel, les acheteurs potentiels étant souvent des acheteurs stratégiques (producteurs d'assurances), ils disposent en général des Générateurs de Scénarios Economiques qui leur permettant de fournir différents scénarios stochastiques pour les données d'input du modèle de machine learning, à savoir les variables économiques et financières dans notre cas. Il s'agit de la méthode qui apparait la plus fiable car elle est fondée sur différents scénarios stochastiques économiques.

Ne disposant pas de Générateurs de Scénarios Economiques pour notre étude, nous décidons d'explorer un nouvel outil souvent utilisée en machine learning **: Le modèle de séries temporelles Facebook Prophet**. Il s'agit d'un modèle fondé sur l'analyse de séries temporelles et qui peut être employée pour prédire l'évolution future des indices économiques et financiers.

L'exercice prospectif relatif aux scénarios économiques futurs n'étant pas l'objet de notre problématique actuarielle et les assureurs disposant en général de Générateurs de Scénarios Economiques pour ce faire, nous choisissons d'utiliser un **scénario déterministe** pour les variables économiques et financières (données d'input) sur 2008-2024 par souci de simplification de nos calculs.

a. Projection des variables économiques et financières selon une analyse de séries temporelles:

Les analyses des séries temporelles concernent l'étude de données collectées à intervalles réguliers sur une période donnée. Cette approche permet d'identifier des tendances, des cycles, des saisons et d'autres composantes inhérentes aux données qui évoluent au fil du temps. En statistique et en économétrie, les séries temporelles sont utilisées pour modéliser et prévoir le comportement futur de phénomènes basés sur leur comportement passé.

Présentation du modèle de séries temporelles Facebook Prophet et application :

Il y a 3 ans, l'équipe de Facebook Core Data Science team à sorti en open source un puissant outil de prévision pour les séries temporelles appelé Prophet.

Le modèle de séries temporelles Prophet est conçu pour traiter les données de séries temporelles avec une approche qui est à la fois robuste et flexible, facilitant la prévision de tendances avec des composantes saisonnières et des effets de vacances. Prophet est particulièrement adapté aux données ayant de fortes saisons et où les données historiques présentent plusieurs saisons de l'historique.

Prophet décompose la série temporelle en plusieurs composantes : tendance, saisonnalité et jours fériés. Cela permet une modélisation fine qui peut s'adapter à divers types de séries temporelles, y compris celles avec des changements non linéaires dans la tendance, des saisonnalités annuelles, hebdomadaires et quotidiennes, et des effets de vacances ou d'événements spéciaux.

Une des forces de ce modèle est avant tout sa facilité d'utilisation. Il nécessite peu de réglages de la part de l'utilisateur pour produire des résultats de haute qualité, ce qui le rend accessible même pour ceux qui ne sont pas experts en statistiques ou en modélisation de séries temporelles.

Prophet gère bien les données manquantes et est robuste face aux données aberrantes (*outliers*), ce qui le rend pratique pour les séries temporelles réelles qui peuvent être incomplètes ou bruitées. De plus, il peut automatiquement détecter les points de changement dans les données pour ajuster la tendance.

Le modèle permet d'intégrer des variables explicatives supplémentaires pour améliorer les prévisions, ce qui peut être particulièrement utile pour modéliser l'impact d'événements ou d'interventions spécifiques.

Prophet fournit également des outils pour évaluer la qualité des prévisions et comprendre les composantes du modèle, facilitant ainsi l'interprétation et l'ajustement des prévisions.

Ce modèle est une procédure de prévision pour les données de séries temporelles basée sur un modèle additif où les composantes non linéaires sont ajustées aux tendances annuelles, hebdomadaires et quotidiennes, ainsi qu'aux effets des jours fériés.

Prophet utilise un **modèle additif** où différentes composantes sont combinées pour faire des prédictions. La formule générale du modèle est la suivante :

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

- $y(t)$ est la valeur prédite à l'instant t
- $g(t)$ est la fonction de tendance qui modélise les changements non linéaires dans les données au fil du temps
- $s(t)$ est la composante saisonnière qui capture les effets saisonniers périodiques
- $h(t)$ représente les effets des jours fériés ou des événements particuliers
- ε_t est le terme d'erreur

Fonction de Tendance $g(t)$

La tendance $g(t)$ dans le modèle Prophet peut être modélisée comme une fonction linéaire ou logistique. Pour une tendance linéaire, on pourrait utiliser $g(t) = k \cdot t + m$, où k est la pente et m l'ordonnée à l'origine. En cas de tendance logistique, on utiliserait une fonction de croissance logistique limitée par un plafond, typiquement modélisée par :

$$g(t) = \frac{C}{1 + e^{-k(t-m)}}$$

Où

- C est le plafond de croissance, c'est-à-dire la capacité de charge (le maximum de la série temporelle),
- k est le taux de croissance,
- m est le décalage (l'offset de croissance)

Composante Saisonnière $s(t)$

La saisonnalité s(t) est modélisée en utilisant une série de Fourier pour capturer les effets saisonniers à différentes échelles de temps. Elle permet d'ajuster le modèle aux motifs saisonniers annuels, hebdomadaires et quotidiens. Cette saisonnalité $s(t)$ peut être modélisée à l'aide d'une série de fonctions trigonométriques pour capturer les motifs cycliques. Par exemple, pour une saisonnalité annuelle, on peut utiliser des termes comme $a_n\ cos(\frac{2\pi nt}{P}) + b_n\ sin(\frac{2\pi nt}{P})$, où P est la période (e.g., 365.25 pour une année) et n est le nombre de termes utilisés pour modéliser la saisonnalité.

Effets des vacances et des jours fériés $h(t)$

Les effets des jours fériés et des événements spéciaux $h(t)$ sont modélisés en ajoutant un terme indicateur pour chaque jour férié qui affecte la série temporelle, permettant au modèle de tenir compte de leurs impacts ponctuels. Il s'agit d'intégrer des ajustements ponctuels lors des jours spécifiques, modélisés par des indicateurs qui prennent des valeurs non nulles pendant ces fameux jours fériés ou vacances.

Les paramètres du modèle sont ensuite ajustés en utilisant une procédure d'optimisation pour minimiser l'erreur de prédiction sur les données historiques. Une fois le modèle ajusté, il peut être utilisé pour faire des projections futures en calculant les valeurs de $g(t)$, $s(t)$, et $h(t)$ pour les périodes à venir. Prophet fournit un cadre flexible pour modéliser et prévoir des séries temporelles, en intégrant de manière explicite les effets de tendance, saisonnalité et événements particuliers, ce qui le rend particulièrement adapté pour les applications où ces facteurs jouent un rôle clé.

Les prévisions qui sont basées sur les données historiques et les hypothèses du modèle doivent être interprétées avec prudence, en tenant compte des facteurs externes et des évènements exceptionnels qui pourraient influencer les résultats futurs.

Le modèle Prophet est conçu pour gérer les tendances et la saisonnalité, mais il peut parfois être sensible aux anomalies ou aux pics exceptionnels dans les données. Si les pics ne sont pas correctement ajustés ou exclus comme des événements atypiques, ils peuvent influencer de manière disproportionnée la tendance future prévue par le modèle.

Après avoir identifié et exclu les anomalies en utilisant la méthode de l'écart interquartile (IQR), le nombre de données a été réduit indiquant que certaines valeurs aberrantes ont été exclues. La moyenne de CPI.relvar2yr a diminué légèrement, passant de 0.042 à 0.031, et l'écart-type a également diminué, passant de 0.035 à 0.017. Cela suggère que les données sont désormais plus homogènes et moins dispersées, ce qui devrait améliorer la qualité des prévisions futures.

Une projection plutôt stable alors que les données historiques montrent une tendance stable peut indiquer que le modèle semble avoir capturé la dynamique sous-jacente des données.

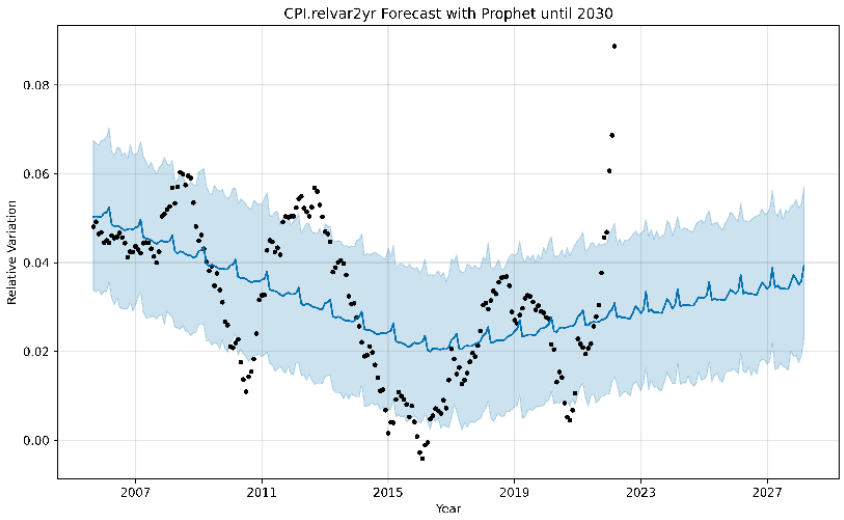


Figure : Projection CPI.relvar2yr selon modèle Facebook Prophet post filtrage (courbe en bleu)

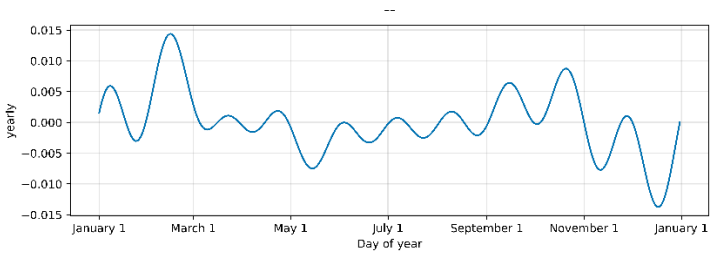


Figure : Composante saisonnalité de la prévision

La désactivation de la saisonnalité hebdomadaire et quotidienne par Prophet se concentre sur les tendances à plus long terme, ce qui est pertinent pour l'analyse du CPI. Cependant nous constatons que le modèle prévoit une inversion de la tendance et le considère comme durable. Cela vient de l'absence d'un niveau suffisant de profondeur de la donnée. Il faudrait en effet entrainer le modèle sur une période historique plus importante pour affiner les projections et retranscrire les cycles économiques.

Nous pourrions utiliser les projections de variables économiques et financières issues du modèle Prophet afin de fournir les données d'input au modèle de machine learning dans les cas où la profondeur de la donnée serait suffisante. Ce n'est pas le cas ici, nous décidons d'utiliser le scénario déterministe suivant par souci de simplification.

b. Projection des variables économiques et financières selon un scénario déterministe :

Pour notre étude nous avons entrainé notre modèle de machine learning sur les données historiques entre 2000 et 2007 et avons considéré que la date de valorisation de l'actif était en 2007. Comme précisé avant, il est préférable que l'acheteur potentiel utilise des Générateurs de Scénarios Economiques, avec comme date de départ la date de valorisation, pour fournir des projections de trajectoires stochastiques de variables économiques et financières qui seront utilisées comme données d'entrée au modèle de machine learning. La date de valorisation pour notre étude étant à fin 2007, nous disposons de données réelles des indices économiques et financiers jusqu'en 2024. Nous avons ainsi la possibilité d'utiliser

ces données réelles relatives jusqu'à aujourd'hui (2024) dans le cadre d'un scénario déterministe de données d'input sur 2008-2024 pour notre modèle de machine learning.

Pour la variable EUidx.relvar2yr (variation relative d'un indice boursier européen sur deux ans), nous avons décidé de retenir la variation relative sur 2 ans de l'indice Euro Stoxx 50.

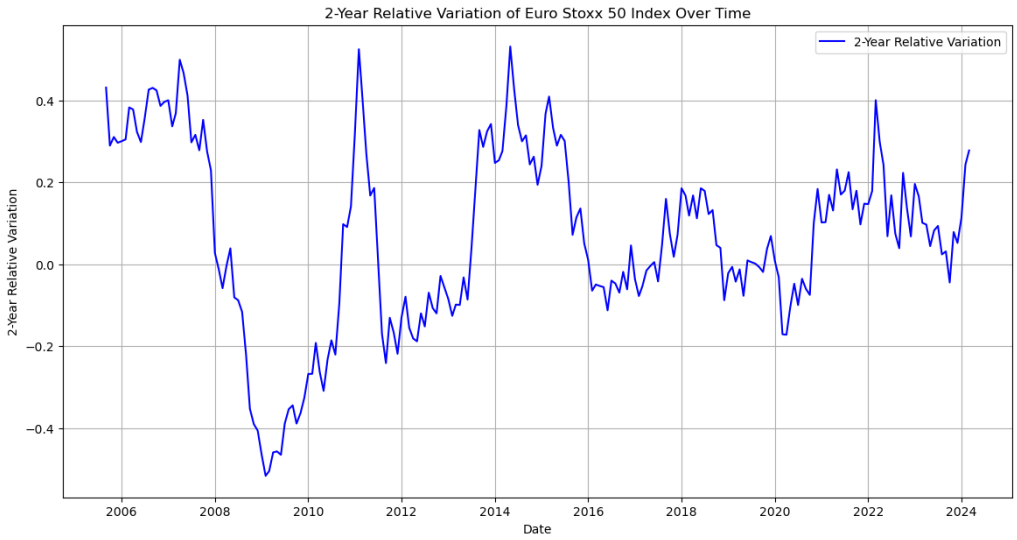


Figure : Variation relative sur 2 ans de l'indice Euro Stoxx 50 sur 2008-2024

Entre 2008 et 2024, plusieurs événements majeurs ont eu un impact significatif sur les marchés financiers mondiaux, y compris l'indice Euro Stoxx 50.
Voici un résumé des événements clés qui ont contribué à la volatilité des marchés financiers, y compris l'indice Euro Stoxx 50, reflétant les réactions des investisseurs aux incertitudes économiques, politiques et mondiales :

- **2008-2009 :** Crise financière mondiale. Déclenchée par l'effondrement du marché immobilier américain, cette crise a entraîné une récession mondiale. Les banques et les institutions financières ont été durement touchées, ce qui a eu un impact négatif sur les indices boursiers mondiaux.
- **2010-2012 :** Crise de la dette dans la zone euro. Plusieurs pays de la zone euro ont été confrontés à des difficultés pour refinancer leur dette souveraine ou pour sauver leurs banques nationales. Cela a entraîné des plans de sauvetage pour des pays comme la Grèce, l'Irlande et le Portugal, et a mis sous pression l'Euro Stoxx 50.
- **2014-2016** : Baisse des prix du pétrole. La chute drastique des prix du pétrole a eu un impact sur les économies mondiales, influençant également les marchés boursiers.
- **2016 :** Brexit. Le référendum au Royaume-Uni a abouti à une décision de quitter l'Union européenne, créant de l'incertitude sur les marchés financiers européens.

- **2020 :** Pandémie de COVID-19. La pandémie mondiale a entraîné des confinements, perturbant l'économie mondiale et les marchés financiers. Les indices boursiers ont initialement chuté avant de se redresser grâce aux mesures de soutien monétaire et fiscal.
- **2021-2024 :** Récupération post-COVID et inflation. La reprise économique rapide, alimentée par des mesures de relance importantes, a entraîné des craintes d'inflation et des ajustements de la politique monétaire par les banques centrales, influençant les marchés boursiers.
- **Tensions géopolitiques et commerciales.** Les tensions entre les États-Unis et la Chine, les sanctions contre la Russie, et d'autres conflits géopolitiques ont également eu un impact sur la confiance des investisseurs et sur les marchés boursiers.

Nous avons aussi récupéré les données relatives aux variations relatives sur 2 ans (CPI_RelVar_2Yr) et 1 mois (CPI_RelVar_1Mth) du Consumer Price Index.

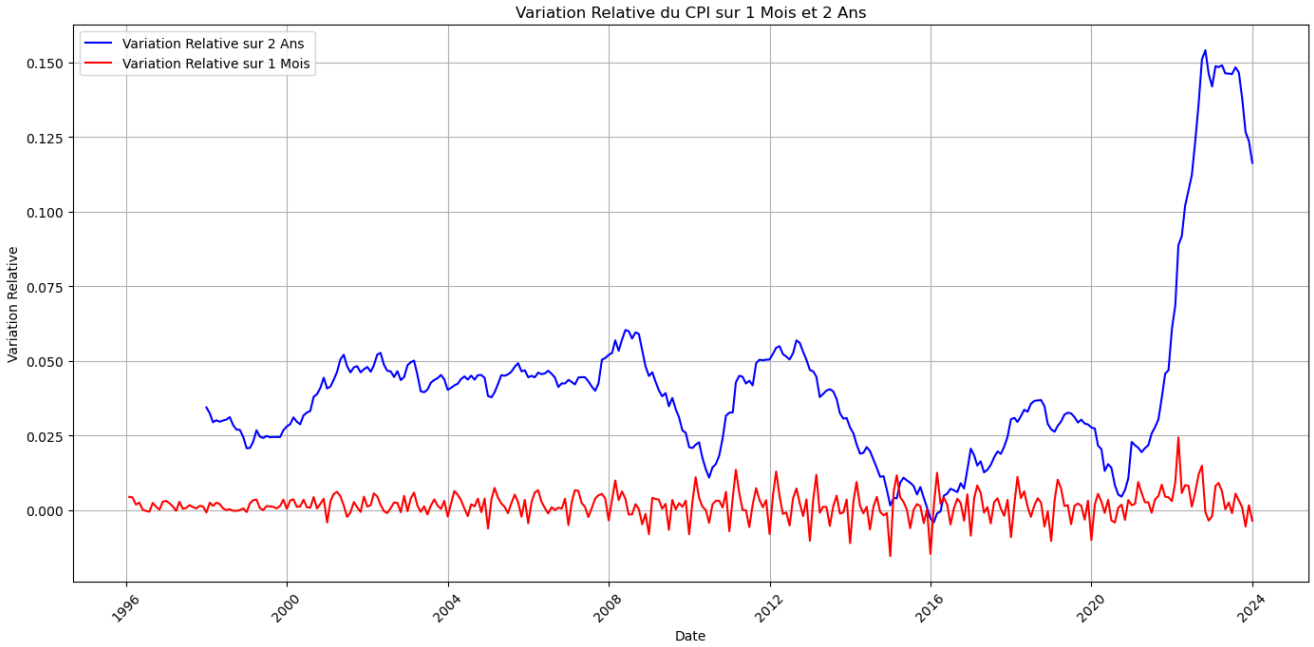


Figure : Variation relative sur 2 ans et 1 mois de l'indice CPI (Consumer Price Index)

La variation relative sur 2 ans du CPI montre une forte volatilité, ce qui indique des changements significatifs dans l'inflation ou la déflation sur cette période. Cette volatilité peut être très pertinente pour notre exercice de prédiction, car elle reflète des mouvements économiques majeurs susceptibles d'influencer le comportement des consommateurs. Par exemple, une forte inflation sur deux ans pourrait réduire le pouvoir d'achat des consommateurs, les rendant plus susceptibles de résilier certains services ou abonnements pour réduire leurs dépenses. Inversement, une période de déflation ou de faible inflation pourrait indiquer une stabilité économique, rendant les consommateurs moins enclins à résilier des services par souci d'économie. Nous pourrons aussi observer si les assurés réagissent bien à ces fortes fluctuations comme le prédit le modèle à travers l'analyse des caractéristiques importantes (*feature importances*).

En revanche, la variation relative par mois semble moins utile pour notre analyse visuelle. Bien que ces informations fournissent des détails sur les fluctuations à court terme de l'inflation, elles peuvent être trop volatiles et sujettes à des variations saisonnières ou des événements ponctuels qui n'ont pas nécessairement un impact significatif sur le comportement à long terme des consommateurs en matière de résiliation.

Plusieurs événements majeurs pourraient avoir impacté l'indice CPI au cours de cette période, contribuant à sa volatilité et soulignant l'importance de prendre en compte ces mouvements économiques majeurs dans notre analyse du risque de résiliation :

- **La crise financière mondiale de 2008-2009**, qui a entraîné une récession globale et des changements significatifs dans les habitudes de consommation.
- **La crise de la dette souveraine européenne autour de 2010-2012**, affectant la confiance des consommateurs et leur pouvoir d'achat dans plusieurs pays de la zone euro.
- **La pandémie de COVID-19 en 2020**, qui a provoqué des confinements à l'échelle mondiale, perturbant l'économie et entraînant des changements radicaux dans les dépenses des consommateurs.
- **Les fluctuations des prix de l'énergie et des matières premières**, influençant directement l'inflation et donc le CPI.

2. Ecoulement du portefeuille en fonction des prédictions de résiliation

A. Ecoulement modélisé par méthode de machine learning

Nous entamons désormais une phase cruciale de notre analyse où nous appliquerons notre modèle développé pour anticiper la manière dont les assurés réagiront aux fluctuations des indicateurs économiques et de marché, tout en prenant en compte l'évolution des caractéristiques propres à chaque assuré.

Cette démarche vise à comprendre en profondeur l'impact de l'environnement économique et financier sur les décisions individuelles de maintien ou de résiliation des contrats d'assurance en épargne.

Plutôt que de nous lancer dans des projections des variables explicatives (par le biais de Générateurs de Scénarios Economiques par exemple), nous avons choisi par souci de simplification et d’efficience de nous appuyer sur des données concrètes et vérifiables qui sont déjà à notre disposition (données réelles 2008-2023).

Ainsi, nous utiliserons les évolutions réelles de deux indicateurs clés : l'indice Euro Stoxx 50, qui reflète la performance des 50 plus grandes entreprises cotées en zone euro, et l'indice des prix à la consommation (Consumer Price Index - CPI), qui mesure l'inflation et la variation du pouvoir d'achat.

Pour ne pas complexifier notre étude, nous avons décidé de limiter l'horizon temporel à une période de 15 ans. Cela a aussi l'avantage de nous permettre d'utiliser un scénario déterministe pour notre étude avec uniquement des données réelles pour nos variables explicatives. Cette durée est suffisamment étendue pour englober plusieurs cycles économiques et permettre une analyse intéressante des tendances, tout en restant assez concise pour éviter les incertitudes trop lointaines et les complexités analytiques excessives.

Pour rappel, voici les variables explicatives relatives aux données économiques et financières les plus importantes que nous utiliserons dans le cadre de nos prédictions :

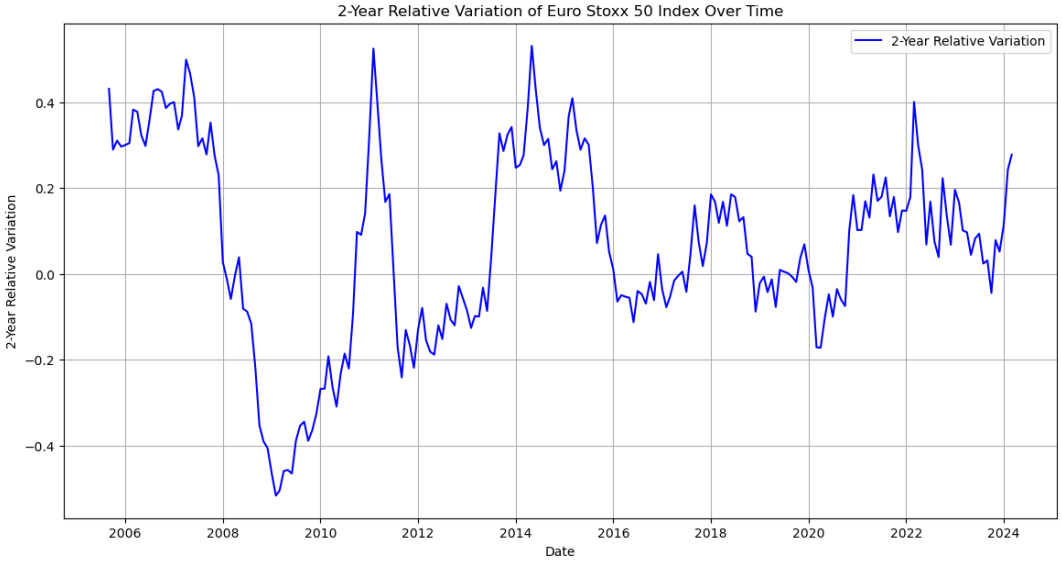


Figure : Variation relative sur 2 ans de l'indice Euro Stoxx 50

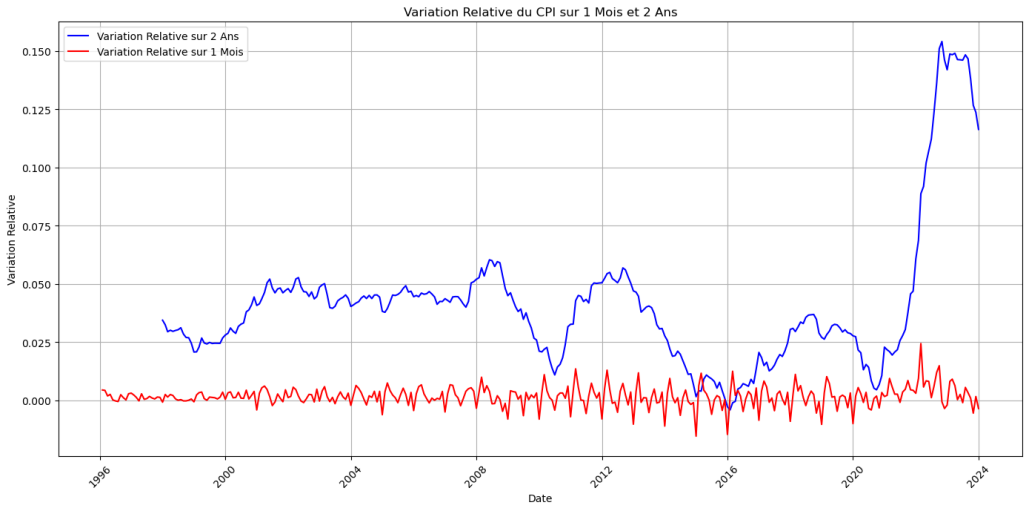


Figure : Variation relative sur 2 ans et 1 mois de l'indice CPI (Consumer Price Index)

Pour estimer les résiliations sur le portefeuille d'assurance durant la période 2008-2023, nous avons utilisé le modèle XGBoost qui a été initialement entraîné sur des données couvrant la période 1999-2007, sélectionné pour ses performances intéressantes.

Lors de l'application de ce modèle sur les années 2008-2023, nous avons procédé à une simulation en faisant varier les variables CPI et EUdixrelvar, conformément aux données économiques réelles observées durant cette période.

Cette approche nous a permis de capturer l'impact potentiel des fluctuations économiques sur les taux de résiliation, en tenant compte des tendances inflationnistes et des dynamiques économiques spécifiques à l'Europe. Au-delà des variables intrinsèques de la clientèle (ancienneté du contrat, âge de la clientèle, etc.), nous avons fait l'hypothèse de ne pas faire varier les autres variables économiques, financières et sociétales.

Voici l'écoulement du portefeuille issu de la prédiction des résiliations du modèle XGBoost sur la base de notre scénario déterministe :

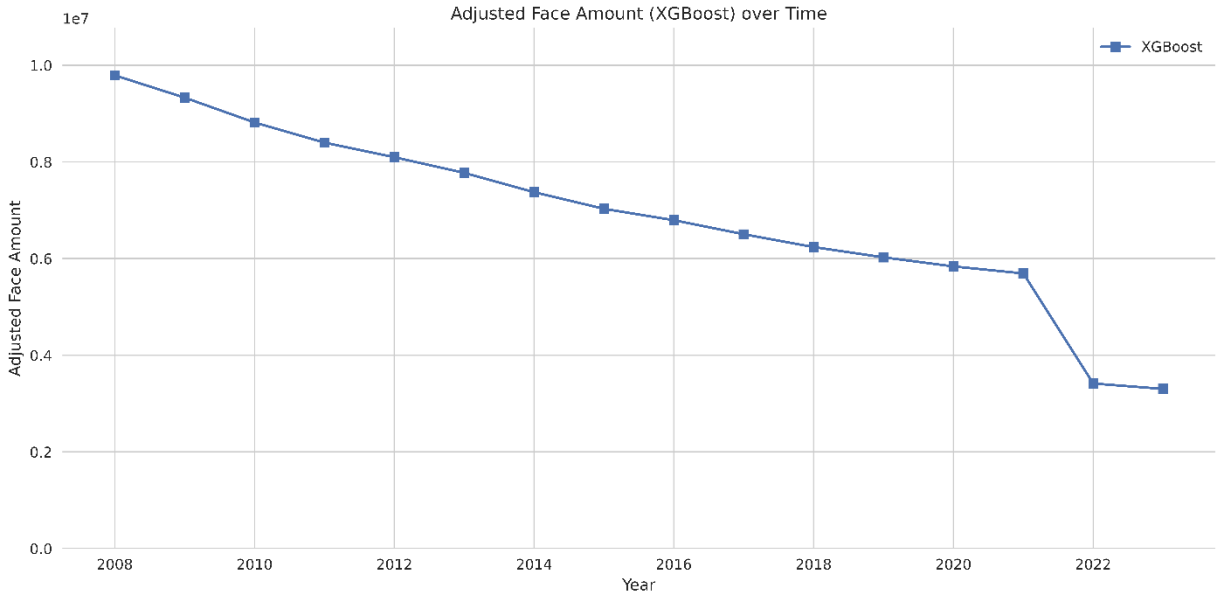


Figure : Evolution du portefeuille selon les résiliations estimées par le modèle XGBoost entrainé

Nous observons que la hausse conséquente des résiliations entre 2020 et 2022 semble suivre le pic de l'indice relative à la variation relative sur 2 ans du CPI.

Une augmentation de l'indice CPI sur deux ans (+14% sur 2020-22) reflète des changements significatifs dans le coût de la vie, ce qui peut entraîner une pression financière accrue sur les ménages. Dans ce contexte, les consommateurs peuvent être amenés à réévaluer leurs dépenses discrétionnaires et leur capacité d'épargne pour s'adapter à l'augmentation des coûts des biens et services essentiels.

Le modèle semble donc bien avoir capté l'effet de l'inflation dans la décision des consommateurs de résilier leurs polices d'assurance.

Le modèle XGBoost a prédit qu'un peu plus de 60% des polices ont résilié sur la période de 15 ans, ce qui représente un taux de résiliation moyen de 7% par an, plutôt en phase avec le taux de résiliation constaté sur l'historique de 6,5% par an.

Ce taux de résiliation peut sembler faible, surtout au regard du pic de croissance. Cela indique toutefois que le calibrage du risque de résiliation, influencé par (i) les fluctuations du marché et du CPI et (ii) l'évolution de l'âge et de l'ancienneté des contrats, nécessite néanmoins une profondeur adéquate de données historiques pour un meilleur calibrage. Avec une profondeur de seulement 7 ans, le modèle ne peut pas bien capter le risque de résiliation pour les anciennetés supérieures à 7 ans.

A titre illustratif et dans un but analytique visuel, nous avons observé les prédictions des résiliations du modèle XGBoost sans faire varier l'ancienneté des contrats et l'âge des assurés, afin de n'observer que l'effet des fluctuations des variables économiques et financières.

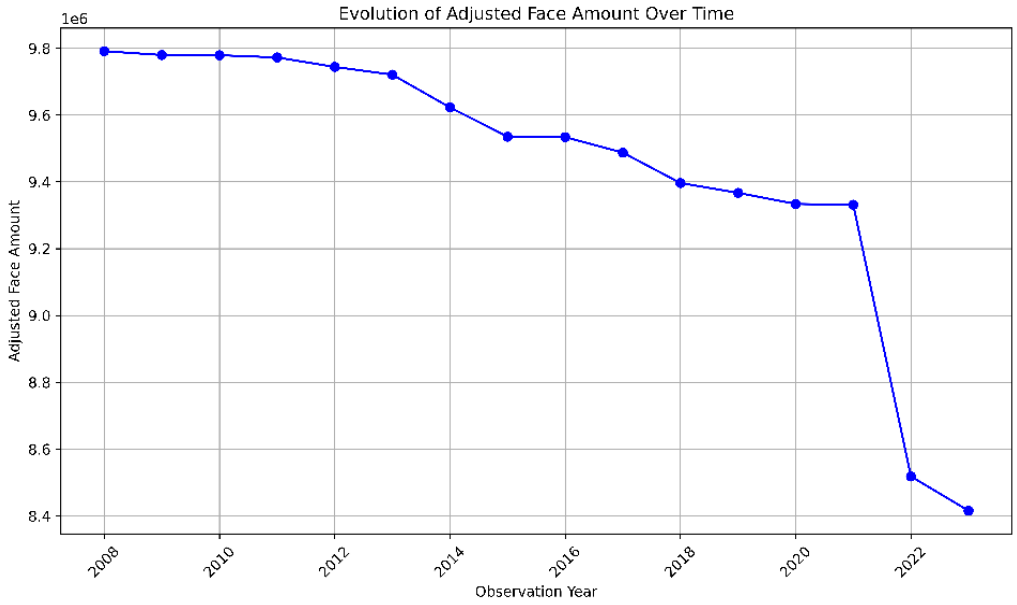


Figure : Evolution du portefeuille selon les résiliations estimées par le modèle XGBoost (sans faire varier l'ancienneté des contrats et l'âge des assurés)

Cette vision permet de constater l'effet sur les résiliations du pic d'inflation via la hausse de l'indice CPI dans la période du covid. Nous observons ainsi que les fluctuations des indicateurs économiques et financiers expliquent 14% des résiliations sur les 60% de résiliations observées avec l'ensemble des paramètres.

B. Comparaison des écoulements selon un taux de résiliation historique vs. issu du modèle de machine learning (XGBoost)

Contrairement aux contrats en Euro, le régulateur n'impose pas aux assureurs de prévoir une composante conjoncturelle dans le cadre des calculs des Best Estimates sur le portefeuille de contrats en UC. Celui-ci n'est donc en général pas considéré car cela aurait un effet pénalisant sur les fonds propres prudentiels.

Dans les méthodes traditionnelles de calibrage du risque de résiliation sur les contrats en UC, il est ainsi usuel de reprendre les résiliations constatées sur l'historique selon les caractéristiques intrinsèques (ancienneté, âge, etc.) des contrats du portefeuille. Dans notre cas, nous constatons un taux de résiliation historique moyen de l'ordre de 6,5% par an.

A des fins de comparaison, nous avons appliqué ce niveau de résiliation historique selon l'ancienneté et l'âge sur le portefeuille étudié (appélé ici la méthode traditionnelle) et nous avons comparé avec le niveau de résiliation issu du modèle de machine learning XGBoost, lui aussi calibré sur l'historique mais intégrant le comportement des assurés face aux fluctuations des indicateurs financiers et économiques.

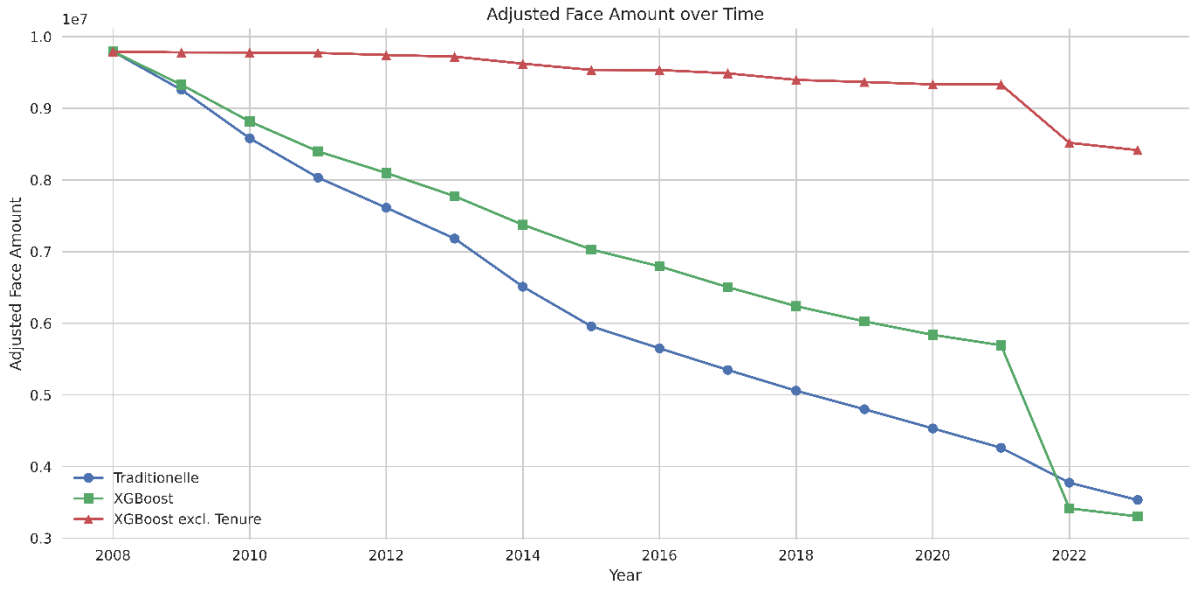


Figure : Ecoulement du portefeuille - Méthode Traditionnelle vs. Méthode de Machine Learning XGBoost (et à titre illustratif XGBoost excl. ancienneté)

La comparaison des écoulements montre que la méthode traditionnelle, qui dépend de l'évolution intrinsèque des caractéristiques du portefeuille comme l'ancienneté et l'âge des assurés, présente un écoulement plutôt stable et quasiment linéaire.

En revanche, la méthode XGBoost met en lumière une chute importante de contrats en réponse au pic d'inflation, un facteur non pris en compte par la méthode traditionnelle. La méthode traditionnelle ne

capte effectivement pas l'effet du pic inflationniste sur le comportement des assurés face à leur optionalité de résiliation.

En définitive, les résultats sont sensiblement les mêmes : la méthode traditionnelle présente 64% de résiliations pendant les 15 années de simulation, tandis que la méthode par machine learning (XGBoost) estime 66% de résiliations, dont 14% des résiliations s'expliquent par le comportement des assurés face au pic inflationniste pendant la période du Covid. Ainsi, nous pourrions déduire par un raccourci que la méthode d'écoulement par machine learning prévoirait environ 50-55% de résiliations sans l'effet inflationniste.

L'explication de ce décalage est délicate, étant donné l'aspect « boite noire » des modèles de machine learning. Cependant la faible profondeur des données sur lequel le modèle XGBoost a été entrainé pourrait expliquer cela : le modèle de machine learning pourrait ne pas avoir capté totalement l'importance de l'effet de l'ancienneté du contrat ou de l'âge de l'assuré face aux autres variables explicatives observées sur la période d'entrainement dont la profondeur n'est que de 7 ans.

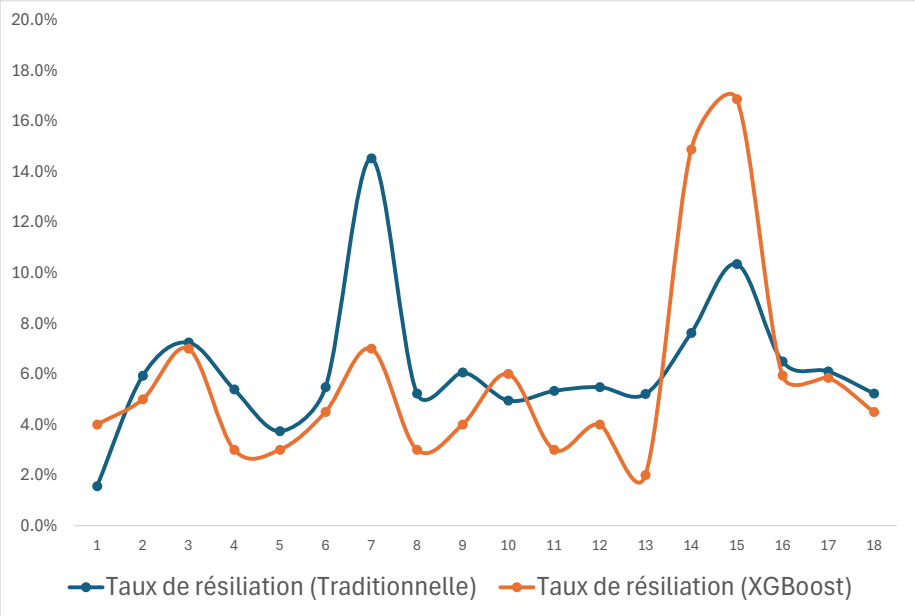


Figure : Taux de résiliation par ancienneté - Méthode Traditionnelle vs. Méthode de Machine Learning XGBoost

Les profils de résiliation par ancienneté montrent en effet un niveau de résiliation pour le modèle XGBoost sensiblement en dessous de la méthode traditionnelle, ne prenant pas en compte le pic de résiliations entre 14 et 16 ans d'ancienneté sous le modèle XGBoost provenant certainement du pic d'inflation du covid.

Le profil de résiliation par ancienneté met en lumière 3 principaux niveaux d'ancienneté pendant lesquels le risque de résiliation serait plus élevé :

- Aux alentours de 3 ans d'ancienneté : surement relatif aux assurés ayant placé leur épargne selon une vision court/moyen terme et qui souhaitent retirer leur épargne pour d'autres projets.
- Aux alentours de 7 ans d'ancienneté : surement lié à un avantage fiscal que bénéficierait le contrat d'assurance vie à partir de 7-8 ans.
- Aux alentours de 15 ans d'ancienneté : surement lié ici à un effet d'échelle avec un portefeuille de contrat résiduel fortement amorti après toutes ces années et un volume de sorties de contrat qui reste stable.

Les comparaisons entre la méthode d'écoulement par machine learning vs. la méthode traditionnelle fondée sur un taux de résiliation historique montre que :

- La méthode de machine learning évalue à la fois le risque de résiliation structurel (ancienneté du contrat, âge de l'assuré, genre, etc.) mais aussi le risque de résiliation conjoncturel lié au comportement des assurés face aux fluctuations de l'environnement économique et financier. L'avantage du machine learning est de capter la composante conjoncturelle qui n'est souvent pas intégrée dans les lois de résiliations des assureurs (car pénalisante dans les calculs d'engagements).
- L'exploitation d'un jeu de données plus profonds en données historiques, comme avec des souscriptions antérieures à 2000 dans notre cas, permettrait une modélisation plus précise par les algorithmes d'apprentissage automatique car cela enrichirait significativement la compréhension des résiliations structurelles du portefeuille (âge, ancienneté, etc.) qui ne sont que faiblement captées par notre modèle ici du fait de la faible profondeur des données. Lors d'une transaction, disposer d'une telle richesse d'informations enrichirait indéniablement l'évaluation de la cible en apportant une contribution complémentaire à la détermination de sa valeur.

3. Les perspectives de valorisation du portefeuille

Sur la base des projections de l'évolution de l'encours du portefeuille, une valorisation serait ensuite possible en appliquant un niveau de chargement sur l'encours afin d'évaluer la marge future intrinsèque du portefeuille.

N'ayant pas à notre disposition les informations relatives aux marges nettes assureurs, nous ne pouvons réaliser l'application numérique sur notre portefeuille projeté. Néanmoins, dans le cadre de la démarche présentée dans cette étude, nous parcourrons les différentes méthodes de valorisation usuelles dans le cadre de transactions assurantielles qui sont les plus pertinentes et davantage conformes aux pratiques

de marché lorsqu'il s'agit de valoriser des cibles spécifiquement en assurance vie épargne, et nous ferons un focus sur la méthode « Embedded Value / Appraisal Value ».

A. Les différentes méthodes de valorisation

Il existe différentes méthodes de valorisation d'une compagnie d'assurance :

- **Comparables de valorisation (ou Multiples de valorisation) :**
  - **Multiples transactionnels :** cette méthode consiste à appliquer les multiples de transactions antérieures concernant des compagnies d'assurance, sur les données de la cible pour laquelle nous cherchons la valorisation. La valorisation s'effectue en utilisant les multiples moyens de ces transactions, basés principalement sur le résultat net pour les actifs en France, mais aussi sur la valeur comptable des capitaux propres. En assurance vie épargne, il est courant d'utiliser les multiples P/UT1 (Unrestricted Tier 1) ou P/EV (Embedded Value, expliqué par la suite).
  - **Multiples boursiers** : Cette méthode repose sur l'utilisation des multiples de valorisation observés sur les marchés boursiers pour des compagnies d'assurance similaires à celle en évaluation. Elle implique l'application des ratios couramment utilisés tels que le P/E (Prix sur Bénéfice), le P/BV (Prix sur Valeur Comptable) ou le rendement du dividende, en se basant sur les données actuelles du marché pour des entreprises comparables. Cette approche permet d'obtenir une estimation de la valeur de l'entreprise cible en la comparant à ses pairs cotés en bourse, offrant ainsi une perspective de valorisation ancrée dans la réalité du marché.

- **Modèle des dividendes actualisés (Dividend Discount Model, DDM)** : Ce modèle offre une vue plus précise de la capacité future de la cible à verser des dividendes à ses actionnaires, basée sur le plan d'affaires de l'entreprise. La capacité de distribution de dividendes est déduite par rapport aux projections de résultats de l'entité à valoriser et introduit la prise en compte du besoin d'immobilisation en capital par rapport à un ratio de solvabilité Solvabilité II cible, généralement dans la fourchette de 150 % à 175 %. Les hypothèses peuvent être ajustées pour observer l'impact potentiel sur la valorisation de scénarios de prix ou de marché défavorables. Cette méthode nécessite de réaliser un business plan crédible de la cible et de prendre en compte les divers éléments de contexte et d'activité de la cible. Cette méthode est très utilisée dans le milieu de l'assurance non vie (branche courte) mais relativement moins dans le milieu de l'assurance vie (branche plus longue).

$$P_0 = \frac{D_1}{(1+r)^1} + \frac{D_2}{(1+r)^2} + \cdots + \frac{D_n}{(1+r)^n}$$

$$= \sum_{n=1}^{\infty} \frac{D_n}{(1+r)^n}$$

Où

P_0 est la valeur des actions de la compagnie

D_n représente les dividendes attendus au titre de chacune des périodes n

r représente le taux d'actualisation qui reflète la rentabilité attendue par un actionnaire pour cette entité. Pour définir ce taux, il est usuel d'utiliser les méthodes de (i) CAPM (Capital Asset Pricing Model) sans coût de la dette afin d'obtenir le coût des fonds propres, et (ii) de Damodaran pour ajuster ce taux selon des primes de risques spécifiques comme entre autres la taille de l'entreprise et le marché dans lequel elle opère.

- **EV (Embedded Value) & Appraisal Value :** Ces deux méthodes consistent en une évaluation économique de la compagnie basée sur la somme des éléments suivants (l'Appraisal Value étant égale à la EV en prenant en compte en plus la valeur des affaires nouvelles (NBV) :
    - ANAV (Adjusted Net Asset Value) ou ANR en français (Actif Net Réévalué) : valeur de l'actif net incluant des ajustements cohérents avec le marché (par exemple, exclusion des incorporels et du goodwill),
    - VIF (Value In Force) : valeur actualisée des profits futurs (nets du coût du capital et des risques non financiers) liés aux polices en vigueur à la date de valorisation,
    - Valeur des affaires nouvelles / New business Value (New business Value) : actualisation des prévisions de valeur de nouveau business.

    Cette méthode nécessite des ajustements basés selon les flux de trésorerie (GAAP français vs TEV vs. MCEV vs. Solvabilité II) pris en compte dans la valorisation. Il est nécessaire de prévoir aussi des analyses de sensibilités de la VIF, de la NBV, des contraintes de capital et des hypothèses d'exigences de solvabilité
    Cette méthode de valorisation est très utilisée en assurance vie car elle offre une mesure qui reflète fidèlement la nature à long terme des contrats d'assurance vie. Elle prend en compte la valeur actuelle des profits futurs attendus des contrats en cours, ce qui correspond étroitement à la manière dont les compagnies d'assurance vie génèrent de la valeur (métier de stock).

Dans le cadre de cette étude, l'érosion du portefeuille permettrait d'obtenir les marges futures qui alimenteraient le compte de résultat de la cible. En appliquant le modèle de machine learning sur le portefeuille en intégrant de manière dynamique les affaires nouvelles sur l'horizon de projection, il serait possible d'obtenir une vision de l'évolution de l'encours dans le temps. Cela permettrait à la fois de réaliser les projections du business plan afin de pouvoir valoriser la cible sur la base de la méthode DDM. Aussi, cette vision de l'évolution de l'encours de la clientèle sur un horizon de projection lointain

permettrait de réaliser une valorisation de la cible selon la méthode EV (Embedded Value) ou Appraisal Value.

Nous allons nous concentrer sur les méthodes relatives à l'EV ou Apparaisal value pour la suite de cette étude au regard de leur caractère très adapté pour la valorisation d'un portefeuille épargne en assurance vie.

B. Focus sur la méthode EV / Appraisal Value

Le bilan d'une société offre une vue instantanée de son patrimoine à un moment précis. Les capitaux propres représentent la richesse cumulée de l'entreprise, accumulée à partir de l'apport initial des actionnaires et des bénéfices réalisés depuis sa création jusqu'à la date d'analyse. Les contrats détenus par l'assureur vont aussi produire des bénéfices à l'avenir, lesquels ne sont pas inclus dans ses capitaux propres actuels. Pour évaluer la valeur d'une compagnie d'assurance, il est nécessaire de considérer à la fois la richesse passée et future.

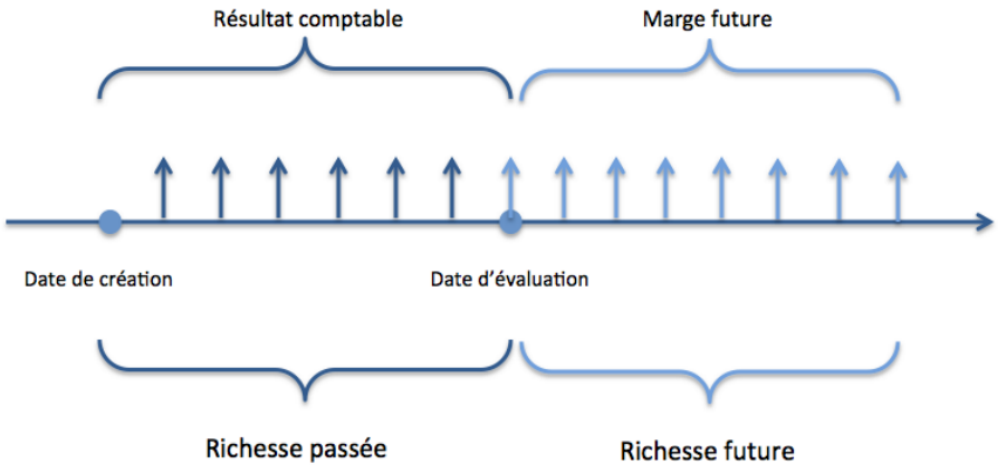


Figure : Richesse d'une compagnie d'assurance vie

Dans ce contexte, nous introduirons un indicateur de richesse spécifique : l'Embedded Value.

Depuis sa création à la fin des années 80, l'Embedded Value sert d'instrument pour évaluer la performance financière des compagnies d'assurance vie, offrant aux actionnaires des insights précieux. Cet indicateur est devenu, ces dernières années, un levier essentiel pour le pilotage de la performance de l'entreprise et joue un rôle crucial dans sa valorisation sur le marché boursier. Toutefois, la baisse des cours de bourse à la fin des années 90 a révélé les limites de l'Embedded Value déterministe, qui tend à sous-évaluer les risques liés à l'activité des assureurs, comme les taux minimums garantis souvent octroyés sans frais aux assurés. Les méthodes de calcul de l'Embedded Value sont régies par les directives du CFO Forum.

Cette association compte vingt-et-une entreprises membres, incluant des acteurs majeurs tels qu'Allianz, Aviva, Axa, BNP Cardif, Generali, entre autres. Les membres du CFO Forum jouent un rôle significatif dans le secteur de l'assurance en Europe.

Le CFO Forum vise à établir des normes pour garantir la cohérence, la normalisation des informations annexes et renforcer la fiabilité des Embedded Value publiées en Europe, offrant ainsi un cadre d'évaluation pour l'Embedded Value tout en permettant aux compagnies d'assurance une certaine flexibilité dans la détermination des hypothèses, l'ajustement des principes et l'application de divers modèles internes. L'embedded value est définie comme étant la « *valeur des intérêts des actionnaires dans les revenus distribuables issus des actifs alloués au business couvert après prise en compte de l'ensemble des risques liés au business couvert* » (principe 3 du CFO Forum).

L'EV permet d'évaluer économiquement une compagnie en sommant l'ANR (ou ANAV - valeur nette de l'actif ajustée du marché, ajusté des plus ou moins-values latentes sur placements et les charges d'impôts y afférente et excluant les incorporels et le goodwill), et la VIF (valeur des profits futurs des contrats existants en portefeuille).

$$EV = ANAV + VIF$$

L'Appraisal value consiste à ajouter la valeur des nouvelles affaires (NBV New Business Value, prévision de la valeur du nouveau business) à l'EV :

$$Appraisal\ Value = ANAV + VIF + NBV$$

Dans un souci de cohérence, les principes de valorisation de la NBV suivent ceux de la VIF.

La VIF (Value In Force) représente la valeur de l'activité d'assurance en cours. Elle est définie comme la valeur actuelle des bénéfices futurs générés par le portefeuille d'assurance, obtenue par projection et ajustée après déduction des impôts correspondants, du coût de l'optionalité des assurés et du coût d'immobilisation du capital. Afin de valoriser l'In Force, trois méthodes de valorisation sont couramment adoptées, se distinguant par leur approche de projection de l'actif (soit déterministe, soit stochastique) et par la manière dont elles évaluent les risques financiers et non financiers :

- La Traditional Embedded Value (TEV)
- L'European Embedded Value (EEV)
- La Market Consistent Embedded Value (MCEV)

La Traditional Embedded Value (TEV) était autrefois le standard de publication de l'Embedded Value (EV), utilisant des projections déterministes des actifs et passifs et incluant le coût des risques financiers et non financiers à travers une prime de risque. En 2004, le CFO Forum a introduit les EEV Principles, établissant l'European Embedded Value (EEV) comme le nouveau standard. Cette approche moderne,

qui repose sur une évaluation stochastique des actifs, prend explicitement en compte le coût des options et garanties financières pour l'assureur.

Ainsi, l'introduction de l'EEV par le CFO Forum a amélioré la pertinence et la comparabilité des Embedded Value publiées, tout en laissant une marge de manœuvre pour le choix des hypothèses et méthodes de calcul. En juin 2008, avec les MCEV Principles, le CFO Forum a standardisé l'approche la plus courante parmi les assureurs publiant une EEV, établissant la Market Consistent Embedded Value (MCEV) comme le nouveau standard de publication dès le 31 décembre 2009. Ces principes ont pour but d'unifier les pratiques diversifiées dans la publication des EV, en insistant sur la cohérence avec les données du marché. Cela nécessite néanmoins un effort significatif dans le calibrage des hypothèses et l'application de modèles de calcul avancés et dynamiques.

La TEV et l'EEV diffèrent de la MCEV dans leur approche de valorisation des risques liés au portefeuille, la MCEV étant la seule à valoriser de façon explicite et par composante les risques affectant l'entreprise. La TEV se valorise la VIF via l'actualisation des flux de trésorerie futurs actualisé via un taux incluant une marge de risque, diminué du coût d'immobilisation/de portage du capital (CoC) qui représente le sous-rendement net pour l'actionnaire lié à l'immobilisation de fonds propres liée à la réglementation :

$$\text{CoC} = \text{T}_{\text{SCR}}\% \times SCR_0 + \text{T}_{\text{SCR}}\% \sum_{\text{t}=1}^{N} \frac{(SCR_t - SCR_{t-1}) - SCR_t \times R_t \times (1 - Taux\ IS)}{(1 + T_{\text{t}})^t}$$

- R_t = rendement des fonds propres en année de projection t
- T_t = taux d'actualisation correspondant au rendement attendu par l'actionnaire
- $\text{T}_{\text{SCR}}\%$ représente le taux de marge de solvabilité minimum cible que devrait avoir l'entité faisant l'objet de la transaction. Dans le cadre de transaction, celui-ci est en général le seuil de marge de solvabilité que le potentiel acquéreur s'impose dans le cadre de la bonne gestion de ses entités d'assurance. Celui-ci est en général aux alentours de 130%-150%

La TEV repose sur une approche de projection déterministe où les tables utilisées pour simuler l'évolution du passif sont fixées selon la politique en vigueur de l'entreprise et respectent les seuils réglementaires minimums. Cette méthode ne considère pas les interactions entre les actifs et le passif. Les projections des actifs sont également réalisées de manière déterministe, généralement en utilisant le taux sans risque auquel s'ajoute le spread découlant de la prime de risque (approche monde réel).

La TEV est aujourd'hui appréciée pour sa simplicité de mise en œuvre et s'utilise souvent comme méthode additionnelle à la MCEV pour confirmer les résultats. Cette méthode comporte néanmoins deux limites principales :

- En tant que modèle déterministe, il valorise uniquement la valeur intrinsèque des risques.
- La valorisation implicite de la valeur temps des risques repose sur un taux d'actualisation margé, dont le calcul peut être complexe et sujet à interprétation. Ce taux est déterminé par une approche descendante (top-down).

Dans l'approche top-down, la prime de risque découle du coût du capital de l'entreprise, tandis que dans l'approche bottom-up, plus conforme à une logique market-consistent, chaque flux futur est actualisé à un taux reflétant le risque spécifique à ce flux. Bien que l'approche top-down ait été initialement préférée pour sa simplicité méthodologique, elle est désormais moins employée que la méthode bottom-up, qui favorise une valorisation davantage market-consistent.

	Market consistent	Traditionnelle
Calcul	Net (de TVOG stochastique)	Déterministe
Rendements	Risque-neutre	Monde réel
Interactions actif-passif	Oui	Oui, limitées
Actualisation	Taux sans risque	Taux avec marge de risque
Coût du capital	FCRC : coûts frictionnels de portage du SCR*	Insuffisance de rendement net d'IS des actifs en couverture du SCR* par rapport à la rémunération attendue de l'actionnaire
Coût des risques non financiers	CNHR : risques non financiers non couvrables	Via la prime de risque

(*) SCR ou (100+x)% du SCR

Figure : Comparaison MCEV et TEV

La Market Consistent Embedded Value (MCEV) est une mesure standardisée qui adopte une approche et des hypothèses alignées avec les normes de Solvabilité 2. Elle utilise les mêmes hypothèses techniques (telles que les frais, les décisions de gestion futures et les lois biométriques) et effectue un calcul stochastique en se basant sur les mêmes hypothèses financières, à l'exception potentielle des participations aux bénéfices. Cette méthode est caractérisée par son calcul net de la Valeur Temps des Options et Garanties (TVOG) de manière stochastique, une actualisation basée sur des taux sans risque (environnement risque neutre) et prend en compte les interactions entre actifs et passifs. Elle vise à refléter les coûts frictionnels de portage du capital requis (FCRC) et inclut un coût pour les risques non financiers non couvrables (CRNHR).

$$MCEV = ANAV + VIF_{MCEV}$$

$$VIF_{MCEV} = PVFP - TVFOG - FRCR - CRNHR$$

La PVFP représente la projection du compte de résultat usuellement entre 50 ans et 70 ans à partir des hypothèses techniques et financières présentées ci-dessous :

Hypothèses techniques

- *Lois de rachats*
- *Lois de mortalité*
- *Lois d'arbitrages*

Paramètres contractuels

- *Taux de chargement*
- *Clauses de participation aux bénéfices*
- *Commissions*
- *Taux garantis*

Hypothèses économiques

- *Produits financiers*
- *Taux d'inflation*

Hypothèses structurelles

- *Taux de frais*
- *Prélèvements fiscaux*

Management actions

- *Politique de distribution de participation aux bénéfices, consommation/dotation de PPE*
- *Politique financière (allocation, achats/ventes, réalisation de PMVL)*

Le FRCR correspond au coût de portage frictionnel du capital requis. Dans un environnement market consistent, ce coût de friction correspond au coût de frottement fiscal ainsi qu'aux frais financiers liés à l'immobilisation de ce capital :

$$\text{FCRC} = \sum_{i=1}^{N} \frac{\text{CR}_i \times \left(TF_i - (TF_i - Frais\ gestion) \times (1 - Taux\ IS)\right)}{(1 + \text{T}_i)^i}$$

Où :

- Le capital requis CR est déterminé comme le capital économique attendu (par exemple 150% du SCR) net des éléments implicites de financement (la VIF nette de Risk Margin et d'IS, les dettes subordonnées et éventuellement le Surplus Fund)
- TF_i = taux forward 1 an observé sur la période i
- T_i = taux sans risque sur la période i

Le CNHR doit être alloué au titre des risques non-couvrables. Il résulte de l'impact asymétrique de risques non-couvrables (risques biométriques et incertitude dans le calibrage des hypothèses de rachats / résiliations ou de frais notamment) ou de la prise en compte de risques non valorisés par ailleurs dans la VIF (risques de défaut, risques opérationnels).

Le CFO Forum ne prescrit aucune méthode spécifique dans l'évaluation du CRNHR. Il indique qu'il devrait être comparable à un coût du capital, évalué selon les principes du projet de réforme solvabilité 2. Un bénéfice de diversification peut être pris en compte lors du calcul du CRNHR.

Cependant, seule la mutualisation entre les risques non couvrables du périmètre modélisé est permise. Aucune diversification entre les risques du périmètre couvert et non couvert n'est permise.

Son calcul s'appuie donc sur une formule de type « Risk Margin » :

$$\text{CNHR} = \%CoC \times \sum_{i=1}^{N} \frac{SCR_{Souscription_i} + SCR_{Op_i}}{(1+\tau_i)^i}$$

Où :

- $\%CoC$ est le coût du capital
- τ_i = taux sans risque sur la période i

In fine la VIF correspond à la valeur économique des contrats en portefeuille. Plus précisément, la VIF est la valeur actuelle des résultats générés par les contrats en stock (PVFP), après déduction :

- du coût des options et garanties financières (les TVFOG permettent de capter le risque financier moyen associé à l'activité),
- du coût des risques non couvrables (le CRNHR qui peut être considérée comme une provision constituée au titre des risques non pris en compte par ailleurs),
- du coût d'immobilisation du capital requis (le FCRC correspondant au coût d'immobilisation des fonds propres exigés par les normes réglementaires et/ou internes).

Le CFO Forum a notamment introduit des principes de valorisation pour la MCEV :

- Les actifs sont valorisés à leur valeur de marché. Les hypothèses économiques doivent permettre de valoriser les passifs de manière market consistent. Selon le 13ème principe des MCEV principles, l'évaluation dans l'univers risque neutre doit retenir comme courbe d'actualisation la courbe des taux swap (comme la courbe EIOPA).

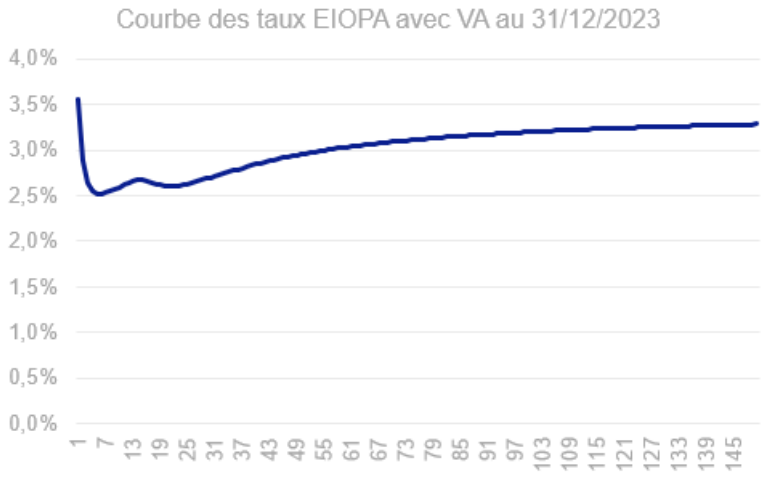


Figure : Courbe EIOPA avec VA au 31/12/2023

- Les éléments incorporels et le goodwill sont exclus de l'Actif Net Réévalué (ANR) car ils sont déjà considérés dans la Value In Force (VIF). Les coûts d'acquisition différés sont également retirés, représentant des dépenses passées.
- Selon le paragraphe 55 des *Basis of Conclusions*, les évaluations se font après déduction de la réassurance, tout en prenant en compte le risque de défaut de celle-ci.
- En accord avec le seizième principe du CFO Forum, les taux de participation aux bénéfices doivent être cohérents avec les pratiques actuelles de la société et du marché, en respectant les réglementations des pays où les contrats sont vendus.
- Les excédents (souvent appelé *Surplus funds*) qui persistent à la fin de la période de projection sont considérés comme appartenant aux actionnaires.
- Les crédits d'impôt et les impôts différés doivent être inclus dans l'ANR, y compris ceux liés à la réserve de capitalisation.
- Pour les éléments non relatifs aux passifs d'assurance inclus dans la MCEV, les normes IFRS sont appliquées.
- Les primes futures incluent dans la MCEV selon le principe 10.2 comprennent les primes périodiques et les versements libres dont les montants sont raisonnablement prévisibles.
- Ces hypothèses, comme définies dans le 11ème principe des MCEV Principles, doivent être rationnelles, estimées au mieux et basées sur des données historiques spécifiques à la compagnie, tout en prenant en compte les tendances futures prévisibles.

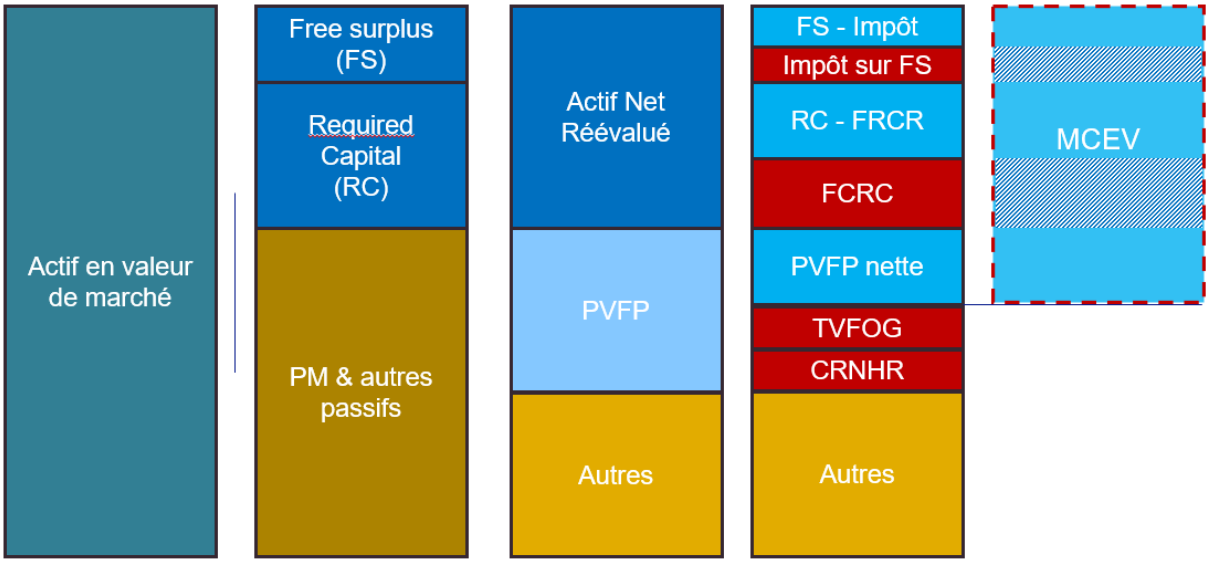


Figure : Composantes de la MCEV

L'utilisation de la MCEV pour la valorisation d'un portefeuille d'assurance vie épargne est ainsi intéressante car elle offre une évaluation conforme au marché, reflétant de manière stochastique la valeur des profits futurs ainsi que le coût des options et garanties financières. Cette méthode permet également une meilleure appréhension des risques et des hypothèses sous-jacentes, en les alignant avec les

pratiques du marché, ce qui conduit à une valorisation transparente et robuste pour les potentiels acquéreurs.

C. Valorisation du portefeuille - Application numérique TEV

Dans le cadre de notre étude, il serait pertinent d'appliquer la méthode d'évaluation MCEV sur la base des scénarios de projections du portefeuille, ainsi que la méthode TEV, sur la base d'un scénario déterministe fondé sur les hypothèses de l'acheteur potentiel, afin de confirmer les résultats. Ne disposant pas de Générateur de Scénarios Economiques pour des travaux stochastiques, nous proposons de ne réaliser que la méthode TEV pour valoriser le portefeuille étudié, sur la base des scénarios déterministes issus des écoulements calculés par machine learning (XGBoost) et par taux de résiliation structurel historique (appelé ici méthode traditionnelle).

Nous sommes partis d'un portefeuille de contrat, issu de données publiques CASdatasets, qui présente les critères relatifs aux assurés et à l'environnement économique et financier, pouvant expliquer le risque de résiliation des contrats. Nous ne disposons par ailleurs pas des taux de marge prélevés sur les contrats de ce portefeuille étudié, ni du détail de rentabilité observable par exemple via les comptes de résultat.

Nous proposons donc des niveaux illustratifs de marges prélevées et de frais afin de réaliser notre exercice de valorisation.

Hypothèses de rentabilité	
Marge financière UC (nette de frais d'acquisition)	0.60%
Marge technique	0.00%
Marge d'arbitrage	0.07%
Frais généraux	-0.30%
Résultat brut	**0.37%**
IS @25,83%	-0.10%
Résultat net	**0.27%**

Nous n'avons pas considéré de marge technique par souci de simplification des calculs. Les 30bps liés aux frais généraux correspondent à un coût normatif illustratif pour la gestion des encours (gestion des sinistres et des contrats) ou au coût de la prestation au travers d'un TSA (Transitional Services Agreement) si seul le portefeuille est concerné par la transaction.

Nous proposons d'actualiser les flux selon un taux d'actualisation, équivalent au coût des fonds propres CoE, calculé comme suivant :

Hypothèses CoE	
Taux sans risque (à la date de valorisation)	2.62%
Prime marché action	8.00%
Beta	0.96x
Prime de risque spécifique	0,0%
CoE (théorique)	**10.30%**

La prime de risque du marché action et le Beta (volatilité sur ce marché) sont issus d'un benchmark de comparables boursiers sur le marché de l'assurance vie.

Parmi les autres hypothèses utilisées pour notre valorisation, nous avons aussi considéré :

Hypothèses CoC & autres	
SCR assureur (% des encours UC)	2%
Marge de solvabilité cible	150%
CoC rate alternatif (RM S2)	5%

Voici ainsi ci-après les projections des flux actualisés de résultat (PVFP) ainsi que du coût du capital (CoC) pour le calcul de la VIF. A cela nous ajoutons l'Actif Net Réévaluée (ANR) pour lequel nous avons considéré égale à 150% du SCR dans le cadre de notre étude. Nous avons considéré de projeter les flux sur 45 ans, en normalisant les flux post 2023.

NB : Dans le cadre de cette étude, la projection des flux est essentiellement influencée par l'écoulement du portefeuille que nous avons obtenu à partir des prédictions par les méthodes de machine learning. Toutefois, dans un exercice réel de valorisation d'un portefeuille d'assurance vie, la projection des flux de marge repose sur de nombreux autres paramètres structurants. Ici, nous avons choisi de nous focaliser sur l'un des aspects clés de la valorisation qui est la prédiction de l'écoulement du portefeuille.

ST

Period		*1*	*2*	*3*		*44*	*45*
Discount factor **traditionnelle)**		*0.91*	*0.83*	*0.75*		*0.02*	*0.01*
VIF (Méthode traditionnelle)	**2008**	**2009**	**2010**	**2011**	**[...]**	**2052**	**2053**
Encours	9 791 377	9 340 100	8 613 671	8 083 904		490 016	457 424
Résultat net		25 632	23 638	22 185		1 345	1 255
PVFP 2009-23	142 975						
PVFP 2024-53	13 186						
PVFP	**156 161**						
SCR	195 828	186 802	172 273	161 678		9 800	9 148
150% SCR	293 741	280 203	258 410	242 517		14 700	13 723
CoC (Approche RM S2)		-14 010	-12 921	-12 126		-735	-686
CoC (actualisé)	**-85 356**						
Marge nette de besoin en capital		11 622	10 718	10 059		610	569
VIF (PVFP - CoC)	**70 805**						
ANR	**293 741**	*hyp. ANR = FP S2*					
Valorisation (ANR + VIF)	**364 546**						

VIF (XGBoost)	**2008**	**2009**	**2010**	**2011**	**[...]**	**2052**	**2053**
Encours	9 791 377	9 329 406	8 815 898	8 399 029		1 253 117	1 211 928
Résultat net		25 603	24 193	23 049		3 439	3 326
PVFP 2009-23	153 261						
PVFP 2024-53	15 463						
PVFP	**168 724**						
SCR	195 828	186 588	176 318	167 981		25 062	24 239
150% SCR	293 741	279 882	264 477	251 971		37 594	36 358
CoC (Approche RM S2)		-13 994	-13 224	-12 599		-1 880	-1 818
CoC (actualisé)	**-92 223**						
Marge nette de besoin en capital		11 608	10 970	10 451		1 559	1 508
VIF (PVFP - CoC)	**76 501**						
ANR	**293 741**	*hyp. ANR = FP S2*					
Valorisation (ANR + VIF)	**370 243**						

Nous avons regroupé l'ensemble des résultats des projections dans le tableau ci-après :

En €		**TEV (méthode traditionnelle)**	**TEV (XGBoost)**
ANR		293 741	
VIF	**PVFP**	154 922	168 724
	CoC	-84 679	-92 223
Valorisation		**363 984**	**370 243**

Dans nos cas d'étude, la valeur du portefeuille de 9,8M€ d'encours de contrats UC vu à fin 2008 est de l'ordre de 360-370K€ selon la méthode d'écoulement des contrats en portefeuille.

Il en ressort une différence de l'ordre de 2% entre les 2 méthodes d'écoulement, ce qui est faible et s'explique par le biais dans la modélisation de l'écoulement du portefeuille de contrats entre les 2 méthodes.

Comme expliqué ci-avant, la faible profondeur des données ne permet pas au modèle de machine learning de capter pleinement le risque de résiliation structurel. Les résiliations modélisées par machine learning sont sensiblement plus faibles qu'en utilisant le taux de résiliation historique. Le pic d'inflation entre 2021 et 2023 vient cependant réhausser le niveau de résiliations dans l'estimation par machine learning, ce qui donne une impression (à tort) que les niveaux d'écoulement sont sensiblement les mêmes entre les 2 méthodes d'écoulement des contrats en portefeuille. Les niveaux d'écoulements étant quasiment similaires à cause de ces biais, cela se reflète ensuite dans la valorisation du portefeuille selon les 2 méthodes explorées, avec seulement 2% d'écart de valeur (les autres hypothèses de valorisation de rentabilité et d'immobilisation de capital étant inchangées).

Nous avons vu au cours de cette étude que les méthodes de machine learning pourraient apporter une contribution additionnelle au calibrage des lois de résiliations dans certains cas. En effet, cette méthode s'avère pertinente dans le cadre de la modélisation du risque de résiliation des contrats en UC dans un contexte transactionnel au regard :

- du niveau d'information limitée concernant le risque de résiliation intrinsèque (structurel & conjoncturel) du portefeuille cible, et
- du biais lié à l'application des lois de résiliation du potentiel repreneur du fait de la non représentativité de la clientèle cible et de l'absence de la composante conjoncturelle de résiliation (car pénalisante et non imposée par le régulateur).

Cette contribution additionnelle qu'apporte le machine learning pourrait ainsi alimenter les hypothèses sous-jacentes sur lesquelles s'appuient l'exercice de valorisation du portefeuille.

La démarche proposée confère ainsi à l'acheteur potentiel une contribution supplémentaire à la meilleure estimation de la valeur intrinsèque du portefeuille de contrats en UC, nonobstant les limites que comporte la modélisation par machine learning et le contexte transactionnel.

Conclusion

La démarche proposée

L'étude s'est concentrée sur l'emploi des modèles de machine learning pour exploiter les données historiques à disposition, représentatifs de la dynamique de résiliation des contrats d'assurance vie en UC et du comportement des assurés du portefeuille face aux conjonctures économiques et financières. Après avoir entrainé les modèles sur la base des données historiques à disposition, nous avons comparé les performances des différents modèles dans le but de sélectionner le modèle le plus pertinent pour l'évaluation du risque de résiliation spécifique au portefeuille qui fait l'objet de la transaction.

Dans un contexte transactionnel, les potentiels acheteurs ne disposent en général pas des détails techniques complets sur les lois de rachat & résiliation de la cible. Il est courant que le potentiel acheteur applique ses propres lois de résiliation pour l'estimation du risque de résiliation du portefeuille cible. Cependant il est fort probable que ces lois ne soient pas représentatives de la clientèle du portefeuille cible, et qu'elles n'intègrent pas la composante conjoncturelle du risque de résiliation car non imposée par le régulateur dans le cadre de ses calculs prudentiels. Néanmoins, les bases de données historiques, telles que les data tapes clients et les AuM tapes, sont en générale transmises et comprennent des données qui sont des informations riches et détaillées sur les contrats des clients, et qui permettent par des méthodes de machine learning de calibrer le risque de résiliation structurel et conjoncturel qui serait spécifiquement représentative du portefeuille concerné.

Implications stratégiques en contexte de transaction

Cette capacité à anticiper les résiliations conjointement structurelles et conjoncturelles, même sans accès direct aux lois de résiliation spécifiques de la cible, se révèle donc intéressante pour affiner les valorisations, et apporter un avantage stratégique lors des remises d'offre ou de négociation dans les transactions. Cela permet aux acquéreurs potentiels de prendre des décisions éclairées, en apportant ainsi un meilleur confort quant à l'appréhension des spécificités des portefeuilles.

Reconnaissance des limitations et perspectives

Bien que prometteuse, l'approche adoptée doit être interprétée avec prudence en raison de la dépendance aux données historiques et des possibles variations dans les tendances de résiliation non capturées par le modèle.

Le niveau d'information mis à disposition constitue un élément clé dans la fiabilité des résultats. Plus l'information disponible est profonde et granulaire, plus les modèles de machines learning seront à même de capter les raisons structurelles et comportementales des assurés quant à leur choix de résiliation. Le risque avec une information trop peu détaillée et profonde, serait d'avoir des modèles qui ne seraient pas capables de capter la majeure partie des raisons impliquant la résiliation et de biaiser

l'importance des critères amenant la résiliation. C'est pourquoi il faut garder en mémoire qu'il s'agit d'un apport complémentaire aux méthodes traditionnelles, et que le niveau de profondeur et de détail de la donnée sont clés pour permettre cette approche.

A l'avenir, l'exploration des modèles prédictifs combinée aux modèles stochastiques sur la base de données adéquates pourraient conduire à un large éventail d'applications dans le domaine de l'assurance et de la banque. Ces méthodes alternatives peuvent offrir un confort supplémentaire en matière de prédiction. En les simplifiant et en facilitant leur mise en œuvre opérationnelle dans des contextes transactionnels, elles peuvent fournir aux assureurs et investisseurs des outils encore plus efficaces et distinctifs pour la prise de décision dans diverses situations. Au niveau des transactions bancaires, ces méthodes de machine learning pourraient aussi apporter de la valeur ajoutée quant au risque de fuite des dépôts, via les bases de données clients (*client tape*), ou bien l'estimation du risque de remboursements de prêt anticipé, via les bases de données des prêts (*loans tape*).

La démarche proposée dans le cadre de ce mémoire permettrait un apport intéressant du machine learning dans l'analyse et la prévision du risque de résiliation des contrats d'assurance vie en unités de compte (UC), en insistant particulièrement sur l'utilité de ces techniques dans un contexte transactionnel.

Ces modèles apportent une alternative aux méthodologies traditionnelles, dont le risque de résiliation conjoncturel n'est pas souvent considéré, capturant la finesse des comportements de résiliation de la clientèle et offrant une précision pertinente pour des évaluations de portefeuille dans le cadre d'une transaction où les lois de résiliations spécifiques ne sont pas toujours transparentes ou accessibles.

Ces nouvelles pratiques ouvrent également de nouveaux horizons pour la modélisation actuarielle, notamment par rapport aux autres enjeux de prévision comme pour le calibrage des lois de décès, d'arbitrage, de rachats partiels & totaux, sur lesquels la modélisation par machine learning pourrait apporter des analyses complémentaires aux méthodes traditionnelles.

Cette démarche vise également à élargir les applications dans le contexte d'IFRS 17, qui introduit notamment des principes de pilotage et de valorisation davantage fondés sur les caractéristiques spécifiques des assureurs liées à leur offre de produits et segmentation de clientèle, sous réserve que les hypothèses sous-jacentes soient justifiées de manière crédible au regard de l'aspect régulé du secteur des assurances.

Elle invite également à une réflexion plus large sur l'utilisation des technologies nouvelles et innovantes dans l'industrie de l'assurance ou de la banque, incitant chercheurs et professionnels à explorer de nouveaux horizons, remettre en question les pratiques traditionnelles et envisager des approches novatrices pour répondre aux enjeux actuels et futur du secteur.

****

Bibliographie

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

[CHOUKARAH, 2019] CHOUKARAH, R. (2019). Prédiction des rachats sur le portefeuille épargne AXA Japon

[Université Bretagne Sud, 2019] Université Bretagne Sud (2019). M1 DSMS - Apprentissage Statistique - *http://thatit.free.fr/COURS/Cours_7.pdf*

[Datasciencetoday] Datasciencetoday.net. Apprentissage avec les arbres de décision - *https://datasciencetoday.net/index.php/fr/machine-learning/109-ml-sup/188-apprentissage-avec-les-arbres-de-decision*

[CAYLA] CAYLA, B. datacorner.fr. Biais & Variance … dilemme ou compromis ? - *https://datacorner.fr/biais-variance/*

[MEHFOOZ (2022)] MEHFOOZ, F. (Kaggle notebook, 2022). Classification with Model Interpretation - *https://www.kaggle.com/code/fahadmehfoooz/classification-with-model-interpretation/notebook*

[ANDRESHG (2021)] ANDRESHG. (Kaggle notebook, 2021). TimeSeries Analysis 📈 A Complete Guide - *https://www.kaggle.com/code/andreshg/timeseries-analysis-a-complete-guide*