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

Par : Madame / Monsieur Qingxia WANG	
Titre du mémoire : Modélisation de résiliation en cas de déséquilibre des classes et mesure de l'élasticité	
Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)	
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.	
Membres présents du jury de la filière : Signature :	Entreprise : Nom : Malakoff Humanis Signature :
	Directeur de mémoire en entreprise Nom : SUN Ao Signature :
Membres présents du jury de l'Institut des Actuaires : 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 du candidat :

Résumé

Dans le domaine de l'assurance santé collectif, le contexte réglementaire actuel génère une concurrence accrue entre assureurs. La perte des clients constitue un vrai problème pour les assureurs en situation de concurrence, parce qu'il est généralement plus coûteux d'acquérir un client que de se battre pour le garder.

La maîtrise des résiliations est d'autant plus importantes que celles-ci s'inscrivent dans un contexte réglementaire en évolution : la résiliation infra-annuelle a été instituée par la loi n° 2019-733 du 14 juillet 2019 [1]. Elle permet à tout agent territorial de résilier et de changer son contrat de complémentaire santé en cours d'année, à l'issue de douze mois de couverture, sans frais.

Ce mémoire a deux objectifs, le premier est de prédire la résiliation d'un client sur un portefeuille donné et de pouvoir trouver des éléments explicatifs de ce comportement. Le deuxième est de modéliser l'élasticité au taux d'indexation individuelle pour répondre aux problématiques d'optimisation tarifaire.

Après une présentation du contexte de l'étude, je décris les principales variables de la base de données construite et le périmètre d'étude. Ensuite, je réalise une analyse statistique. Dans la première partie de modélisation, nous avons utilisé des modèles de classification binaire où l'on cherche à prédire si les clients vont résilier ou non. Les méthodes utilisées sont tirées du Machine Learning : régression logistique, l'arbre de décision, forêt aléatoire, Gradient Boosting et Machine à vecteurs de support. Les hyperparamètres sont optimisés en utilisant l'AUC tirée de la courbe ROC combinée avec un Grid Search et une validation croisée. Le meilleur modèle est choisi à l'aide des matrices de confusions, F1-score et l'AUC. Le meilleur modèle se révèle être la régression logistique. Ce modèle fournit de meilleurs résultats en termes de prédiction sur la base test. Ensuite, l'importance et le rôle des variables sont étudiés, ces variables concernent la cotisation, la prestation. De plus, le diagramme de dépendance partielle est utilisé pour mieux comprendre comment les variables importantes affectent les probabilités d'événements estimées de la résiliation prévue. Dans la deuxième partie de modélisation, nous mesurerons l'élasticité de résiliation au taux d'indexation en utilisant le modèle de régression logistique.

Mots clés : Classes déséquilibrées, Ré-échantillonnage, SMOTE, Forêts aléatoire, Régression Logistique, L'arbre de décision, Forêt aléatoire, Gradient Boosting et Machine à vecteurs de support.

1. La loi n° 2019-733 du 14 juillet 2019 relative au droit de résiliation sans frais de contrats de complémentaire santé permet aux assurés de résilier, après un an de souscription, leur contrat de complémentaire santé, à tout moment, sans frais ni pénalité.

Abstract

In group health insurance, before the introduction of the infra-annual termination law (RIA), policyholders could only terminate their contract at the end of the year by sending a registered letter to the insurance organization at least two months in advance of the expiry date. The infra-annual termination law will give policyholders of group contracts the possibility to terminate their supplementary health insurance contract at any time, once the first year of subscription has elapsed. For equivalent coverage, it, therefore, seems reasonable to assume that policyholders will be tempted to terminate their contract(s) if they consider the prices to be costly compared to what the competition would offer. Insurers are now expecting increased competition. To apply a retention policy and to keep the portfolio balanced. It is important to better understand this risk of cancellation.

This paper has two objectives, the first is to predict the termination by a customer on a given portfolio and to find explanatory elements of this behavior. The second is to model the individual indexing elasticity to answer the problems of tariff optimization.

After a presentation of the context of the study, I describe the main variables of the constructed database and the scope of the study. Then, I perform statistical analysis. In the first part of modeling, we use binary classification models where we try to predict if the customers will cancel or not. The methods used are taken from Machine Learning : logistic regression, decision tree, random forest, Gradient Boosting, and Support Vector Machine. The hyperparameters are optimized using the AUC from the ROC curve combined with Grid Search and cross-validation. The best model is chosen using the confounding matrices, F1-score and AUC. The best model was found to be logistic regression. This model provides better results in terms of prediction on the test basis. Next, the importance and role of the variables are studied, these variables concern the contribution and the benefit. In addition, the partial dependence diagram is used to better understand how the important variables affect the event probabilities of the expected termination. In the second part of the modeling, we will measure the elasticity of termination at the indexation rate using the logistic regression model.

Keywords : Unbalanced classes, Resampling, SMOTE, Random forests, Logistic regression, Decision tree, Gradient Boosting, Support Vector Machine.

Remerciements

Je tiens tout d'abord à remercier mon encadreur Madame Maud Thomas puisqu'elle est non seulement un excellent professeur capable de vulgariser une partie de ses connaissances extensives et de les transmettre à un public parfois néophyte, mais également une directrice généreuse et serviable.

Je voudrais remercier mon tuteur en entreprise Ao SUN pour avoir soutenu et accompagné lors de mon alternance, pour ses conseils durant la rédaction de ce mémoire et pour avoir aiguisé mon intérêt pour le machine learning et l'actuariat.

Je tiens à exprimer toute ma reconnaissance à mon supérieur Pascal BOISSON. Je le remercie de m'avoir encadré, orienté, aidé et conseillé.

J'adresse mes sincères remerciements à mon collègue Renaud RIOUAL, qui a guidé mes réflexions et a accepté de m'aider et de répondre à mes questions durant la rédaction de ce mémoire.

Merci à mon collègue Tanguy RIOUST DE LARGENTAYE d'avoir pensé à me faire découvrir un domaine jusqu'alors pour moi inconnu et qui me passionne aujourd'hui : l'inférence causale et de la Data Science.

Enfin, toute ma reconnaissance va également vers mes collègues de l'équipe ID2, pour leur disponibilité.

Note de synthèse

Contexte et problématique

La révision tarifaire annuelle des contrats est un processus qui intervient le plus souvent une fois par an, à la date d'échéance anniversaire du contrat, afin de garder l'équilibre de portefeuille des assureurs. Une augmentation de prime peut éventuellement conduire les clients à résilier leurs contrats, s'ils peuvent trouver une autre couverture de complémentaire santé qu'ils jugent plus compétitive ou plus intéressante. En outre, la mise en application de la loi RIA, facilitant les démarches de résiliation après un an de souscription, accentue le risque de départ du client. En effet, le départ d'une catégorie de clients conduirait au déséquilibre du portefeuille, nécessitant d'ajuster les tarifs, entraînant possiblement de nouveaux changements de la base de clients (par arrivées et départs de clients). L'objectif devient alors pour l'assureur de bien identifier les profils de risque et adapter la stratégie de tarification.

Dans ce mémoire, nous allons travailler avec des données de résiliation étiquetées et fortement déséquilibrées. Dans un premier temps, les différentes méthodes de classification sont appliquées sur notre base de données déséquilibrée. Ensuite, nous allons vérifier si nous pouvons améliorer les performances des modèles en ré-échantillonnant, c'est-à-dire ré-échantillonner les données pour se rapprocher d'une situation d'équilibre. Pour ce faire, nous expliquerons quatre méthodes de ré-échantillonnage différentes et évaluerons leurs effets sur l'application de prédiction de la résiliation.

Classification avec des classes déséquilibrées

L'ensemble de données de résiliation présentent une distribution déséquilibrée des classes. Une classe déséquilibrée signifie qu'une classe est représentée par un grand nombre (majorité) d'échantillons plus qu'une autre (minorité).

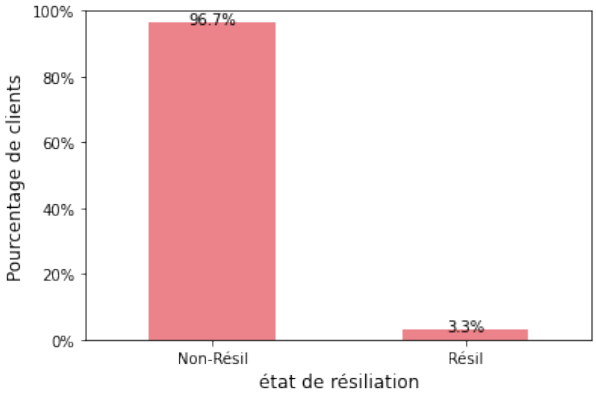

FIGURE 1 – La répartition de nombre de clients en fonction de la variable cible résiliation en 2021

Par exemple, dans notre ensemble de données de recherche, il y a 96.7 % classées comme "Non-Résilié" tan-

dis que 3.3 % sont classées comme "Résilié". L'apprentissage automatique est généralement prédisposé par les données déséquilibrées parce que la plupart des algorithmes standards s'attendent à des distributions de classes équilibrées, ce qui fait que les algorithmes d'apprentissage sont peu performantes pour les données déséquilibrées. De nombreuses applications du monde réel sont critiques pour l'apprentissage des données déséquilibrées, comme le diagnostic médical, la détection des fraudes.

Méthodes de classification

Régression logistique
La régression logistique est une technique de modélisation très couramment utilisée pour décrire la relation existant entre une variable d'intérêt binaire et une ou plusieurs variables explicatives. Dans ce modèle, la réponse suit une distribution binomiale et la fonction logit permet de relier la probabilité d'un résultat positif $(Y = 1)$ à la valeur du prédicteur linéaire.

Arbre de décision
Un arbre de décision est un outil d'aide à la décision. Il emploie une représentation hiérarchique de la structure des données sous forme des séquences de décisions en vue de la prédiction d'un résultat ou d'une classe. La procédure commence par choisir, parmi les variables explicatives, celles qui sont les plus discriminantes pour la variable d'intérêt et qui permettent de séparer les individus en sous-groupes homogènes appelés nœuds. Chaque nœud contient les individus d'une seule classe.

Forêt aléatoire
La forêt aléatoire est un algorithme incontournable en machine learning. Il consiste à améliorer les arbres de décision CART en utilisant du bagging, dans le but de rendre les arbres utilisés plus indépendants (moins corrélés).

Gradient Boosting
Le gradient boosting est une technique d'agrégation de modèle qui permet de transformer les apprenants faibles en apprenants forts.

Machine à vecteurs de support
L'idée de la machine à vecteurs de support est de trouver un hyperplan qui sépare au mieux un jeu de données et maximiser la distance entre ces deux classes.

Méthodes de ré-échantillonnage

Sous-échantillonnage aléatoire
Une technique consiste à éliminer aléatoirement de quelques exemples de la classe majoritaire pour diminuer leur effet sur le modèle. L'inconvénient principal de cette approche est qu'elle peut supprimer des cas importants qui fournissent des informations riches à la différenciation des deux classes.

Sur-échantillonnage aléatoire
Cette technique cherche à tirer de façon aléatoire des points de la classe minoritaire, puis à les dupliquer pour augmenter leur effet sur le modèle. L'inconvénient majeur de cette approche est qu'elle peut conduire à un risque de sur-apprentissage du modèle.

Synthetic Minority Over-sampling Technique (SMOTE)
Cette méthode de sur-échantillonnage se concentre sur la classe minoritaire, qui consiste à générer des données "synthétiques".

SMOTE & TomekLinks
L'approche Smote-TomekLinks repose sur l'idée de créer de larges zones sûres, en appliquant l'approche par Tomek Links(en supprimant les deux points de chaque paire Tomek Link) après avoir sur-échantillonné la base originale avec la technique SMOTE.

Mesures de performance

La matrice de confusion
La matrice de confusion croise la classe réelle des individus avec la classe prédite par le modèle. La matrice de confusion est un tableau spécifique, où chaque ligne représente les instances d'une classe réelle, et chaque colonne les instances d'une classe prédite (ou vice-versa).

	Classe prédite	
Classe réelle	Résilié $Y = 1$	Non - Résilié $Y = 0$
Résilié	VP	FN
Non-Résilié	FP	VN

Accuracy
L'accuracy décrit la quantité relative de données correctement classées dans un ensemble de données. Si $\hat{y}_i$ est la prédiction pour le point de données i et y_i est l'étiquette réelle, le mauvais classement est défini comme suit :

$$accuracyc_n = \frac{1}{n} * \sum_i \left(y_i = \hat{y}_i\right).$$

Précision (Taux de positifs prédits)
La précision est également appelée Positive Prédictive Value. Elle correspond au taux de prédictions correctes parmi les prédictions positives : $\frac{VP}{(VP+FP)}$.

Recall (Taux de vrais positifs)
Le recall est ainsi appelé rappel (sensibilité), true positive rate ou encore hit rate (taux de détection). Il correspond au taux d'individus positifs détectés par le modèle $\frac{VP}{(VP+FN)}$.

F1 score
La F1-score correspond à un compromis entre la précision et le recall.

$$F1 - score = \frac{2 \times \text{ recall } \times \text{ precision}}{\text{recall+precision}}.$$

Modélisation de résiliation en cas de déséquilibre des classes

Les modèles sont obtenus avec le meilleur ensemble de paramètre. Les hyperparamètres ont été ajustés par grid-search. Le grid-search est une méthode d'optimisation qui va nous permettre de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage.

Résultat basé sur la base originale

Une fois les classificateurs formés, leurs performances sont comparées à l'aide de la matrice de confusion pour déterminer la précision, le recall et le f1-score.

TABLE 1 – Récapitulation résultats sur la base déséquilibrée

Modèle	Accuracy	Précision	Recall	Score-F1	ROC AUC
Régression logistique	0.931	0.106	0.163	0.129	0.618
Arbre de décision	0.804	0.058	0.347	0.100	0.572
Forêt aléatoire	0.699	0.056	0.541	0.101	0.635
Gradient Boosting	0.883	0.076	0.245	0.116	0.620
Machine à vecteurs de support	0.877	0.077	0.265	0.119	0.639

Ces méthodes obtiennent les valeurs AUC de 0,62 environ, ce qui est faible. Nous pouvons constater que ceux ci sont plus élevés pour le modèle de machine à vecteurs de support. Selon le score-f1, nous constatons que la capacité de prédiction de nos modèles est mauvaise.

Résultat basé sur la base ré-échantillonnage

Dans cette étude, 70 % des observations ont été utilisées pour l'apprentissage, et les 30 % restants ont été utilisés pour la base de test. Pour rappel, la base d'apprentissage est constituée de 242 entreprises résiliées et de 7 017 entreprises non résiliées. La proportion de cas positifs est de 3,3 %. La distribution de résiliation d'après ré-échantillonnage est les suivants :

TABLE 2 – Répartition des effectifs des bases ré-échantillonnées

	Non-Résilié (0)	Résilié (1)	Total
Base déséquilibrée	7 017 (96.7 %)	242 (3.3 %)	7 259
Sur-échantillonnage	7 017 (50 %)	7 017 (50 %)	14 034
Sous-échantillonnage	242 (50 %)	242 (50 %)	484
SMOTE	7 017 (50 %)	7 017 (50 %)	14 034
SMOTE &TomekLinks	6 995 (50 %)	6 995 (50 %)	13 990

La précision fait référence au pourcentage de clients que le modèle prévoit de résilier et qui ont effectivement résilié. Par conséquent, on peut dire que la précision montre à quel point on peut être certain des prédictions positives. Elle ne dit cependant rien sur les clients qui vont résilier mais que le modèle classe à tort comme

non-résilié. Les modèles produisent une précision assez faible (voir le tableau ci-dessus).

Le recall, quant à lui, s'intéresse aux clients qui vont effectivement résilier, et combien d'entre eux ont été trouvés. Il s'agit également d'une mesure importante pour comprendre comment les modèles peuvent prédire les résiliés.

Comme il est plus coûteux de ne pas prédire un échantillon de résiliation que de prédire faussement un échantillon de non-résiliation, notre intérêt est d'essayer de prédire correctement la classe résiliée. Avec l'utilisation de la technique de ré-échantillonnage, nous voulons augmenter le nombre de vrais positifs et diminuer le nombre de faux négatifs, c'est-à-dire que nous voulons augmenter le recall (rappel).

TABLE 3 – Récapitulation résultats avec sous-échantillonnage

	Recall				
	Base originale	Sur-échantillonnage aléatoire	Sous-échantillonnage aléatoire	Smote	Smote & TomekLinks
Régression logistique	0.163	0.519	0.910	0.038	0.01
Arbre de décision	0.347	0.010	0.903	0.558	0.077
Forêt aléatoire	0.541	0	0.923	0	0
Gradient Boosting	0.245	0.202	0.913	0.019	0.009
Machine à vecteurs de support	0.265	0.250	0.817	0.250	0.26

Le tableau 3 montre que l'utilisation de la technique de sous-échantillonnage permet d'augmenter le nombre de vrais positifs (autrement dit d'augmenter le nombre de classes de résilié correctement prédit). Cependant, pour l'ensemble de données déséquilibré, la forêt aléatoire présente le score le plus bas parmi les méthodes d'ensemble de sur- échantillonnage, mais a montré des scores plus élevés pour l'ensemble de données originaux et les ensembles de données sous-échantillonnés.

Modélisation de l'élasticité de résiliation au taux d'indexation

L'élasticité de la probabilité de résiliation au taux d'indexation correspond à la variation relative de cette probabilité lorsque le taux d'indexation varie d'une quantité infinitésimale. Ceci permet de connaître la réaction des assurés face à une variation de taux d'indexation, en termes de résiliations.

La Figure 2 nous montre la distribution de l'élasticité de 2019 à 2021. On vient de voir qu'il y a une augmentation de l'élasticité de 2020 à 2021. En résumé, l'élasticité par rapport au taux d'indexation est généralement positive et statistiquement significative, mais sa valeur varie de 0 à 1. Ces résultats indiquent que les gens font attention au changement de taux d'indexation auquel ils sont confrontés et à leur propre risque d'avoir besoin de prestations.

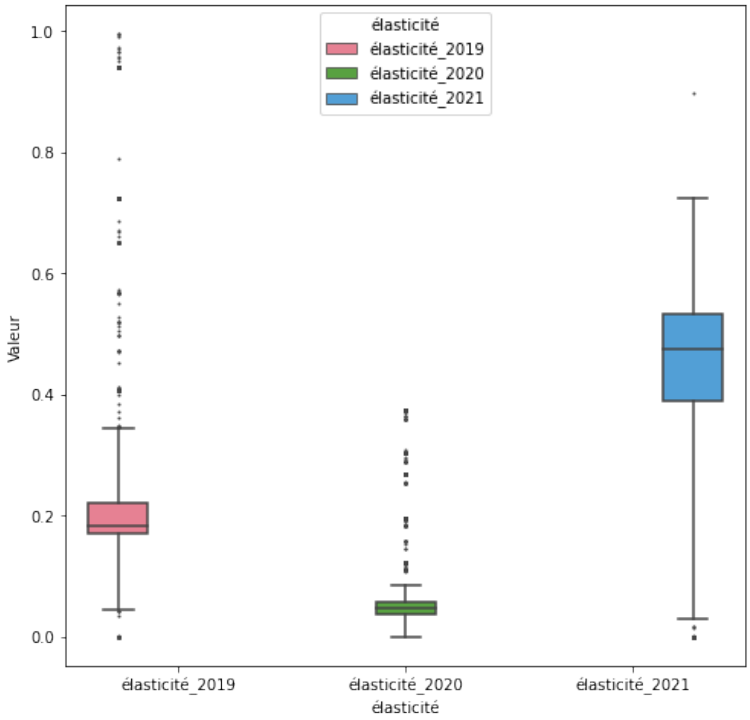

FIGURE 2 – Élasticité moyenne du taux de résiliation au taux d'indexation

Conclusion

Dans ce mémoire, la tâche consiste à identifier les profils de clients qui ont une probabilité de résiliation plus élevée. Une comparaison des différents algorithmes d'apprentissage a permis de conclure, que pour notre jeu de données, les résultats ne sont pas satisfaisants. Afin de renforcer le pouvoir prédictif de nos modèles, nous avons proposé une méthodologie pour améliorer la prédiction de la classe résiliée (minoritaire) en utilisant des techniques de ré-échantillonnage. Nous avons prouvé que les prédictions apportées par l'approche sous-échantillonnage étaient les plus performantes. Ensuite, pour la partie sur l'élasticité, le modèle logistique aide à fournir des coefficients pour la mesure de l'élasticité de résiliation au taux d'indexation. L'étude de l'élasticité nous permet d'optimiser le processus de révision tarifaire en vue d'ajuster mieux le tarif correspondant au risque à date, et également minimiser le risque de résiliation.

Synthesis

Context and problematic

The annual rate review of contracts is a process that takes place once a year, on the anniversary date of the contract, in order to keep the insurers' portfolio balanced. An increase in premium may eventually lead clients to cancel their contracts, if they can find another complementary health insurance coverage that they consider more competitive or more interesting. In addition, the implementation of the RIA law, which facilitates the cancellation process after one year of subscription increases the risk of client departure. Indeed, the departure of a category of customers would lead to an imbalance in the portfolio, necessitating an adjustment of the rates, possibly resulting in new changes in the customer base (through customer arrivals and departures). The objective then becomes for the insurer to properly identify the risk profiles and adapt the pricing strategy.

In this dissertation, we will work with labeled and highly unbalanced termination data. First, the different classification methods are applied on our unbalanced database. Then, we will check if we can improve the performance of the models by resampling, i.e. resampling the data to get closer to an equilibrium situation. To do so, we will explain four different resampling methods and evaluate their effects on the termination prediction application.

Classification with imbalanced classes

The termination data set has an unbalanced distribution of classes. An unbalanced class means that one class is represented by a large number (majority) of samples more than another (minority).

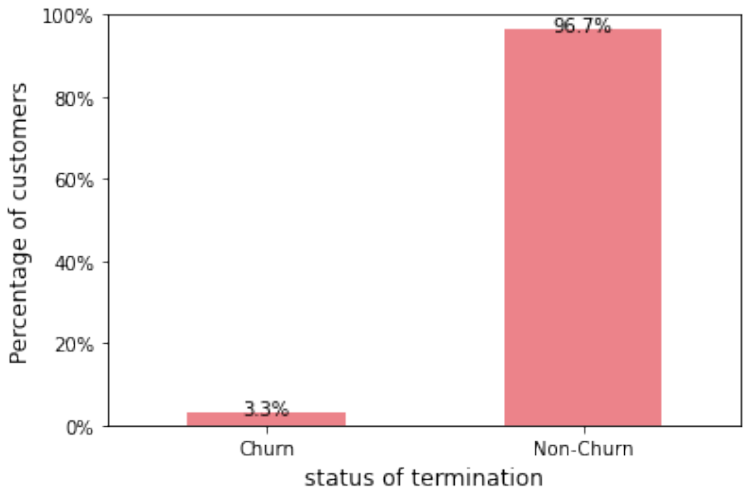

FIGURE 3 – Distribution of the number of customers according to the target variable churn in 2021

For example, in our data set, there are 96.7 % classified as "non-churn" while 3.3 % are classified as "churn". Machine learning is generally predisposed by unbalanced data because most standard algorithms expect balanced class distributions, which makes learning algorithms perform poorly for unbalanced data. Many real-world applications are critical for learning from unbalanced data, such as medical diagnosis and fraud detection.

Classification methods

Logistic regression
Logistic regression is a modeling technique that is very commonly used to describe the relationship between a binary variable of interest and one or more explanatory variables. In this model, the response follows a binomial distribution and the logit function relates the probability of a positive outcome ($Y = 1$) to the value of the linear predictor.

Decision tree
A decision tree is a decision support tool. They use a hierarchical representation of the data structure in the form of decision sequences to predict an outcome or a class. The procedure begins by choosing, among the explanatory variables, those that are the most discriminating for the variable of interest and that allow the separation of the individuals into homogeneous subgroups called nodes. Each node contains the individuals of a single class. .

Random forest
Random forest is an essential algorithm in machine learning. It consists of improving CART decision trees by using bagging, in order to make the trees used more independent (less correlated).

Gradient Boosting
Gradient boosting is again a model aggregation technique that turns weak learners into strong learners.

Support vector machine
A support vector machine is based on the idea of finding a hyperplane that best separates a set of data and maximizes the distance between these two classes.

Re-sampling methods

Random undersampling
One technique is to randomly eliminate a few examples from the majority class to reduce their effect on the model. The main disadvantage of this approach is that it may remove important cases that provide rich information to the differentiation of the two classes.

Random oversampling
This technique seeks to randomly draw points from the minority class, then duplicate them to increase their duplicate them to increase their effect on the model. The major disadvantage of this approach is that it can lead to a risk of over-learning the model.

Synthetic Minority Over-sampling Technique (SMOTE)

This oversampling method focuses on the minority class, which consists of generating "synthetic" data.

SMOTE & TomekLinks
The Smote-TomekLinks approach is based on the idea of creating large safe zones, by applying the the Tomek Links approach (by removing the two points of each Tomek Link - pair) after having oversampled the database.) after oversampling the original base with the SMOTE technique.

Performance measures

Confusion matrix
The confusion matrix crosses the real class of individuals with the class predicted by the model. The confusion matrix is a specific table, where each row represents the instances of a real class, and each column the instances of a predicted class (or vice versa).

	Predicted class	
Actual class	Churn $Y = 1$	Non - Churn $Y = 0$
Churn	VP	FN
Non-Churn	FP	VN

Accuracy
Accuracy describes the relative amount of correctly ranked data in a data set. If $\hat{y}_i$ is the prediction for data point i and y_i is the actual label, misclassification is defined as :

$$accuracyc_n = \frac{1}{n} * \sum_i \left(y_i = \hat{y}_i\right).$$

Precision (Rate of positive predictions)
This indicator represents the proportion of well-predicted positives among all observed positives and is calculated by the observed positives and is calculated by the formula $\frac{VP}{(VP+FP)}$.

Recall (Rate of positive predictions)
Out of the total positive, what percentage are predicted positive. It is the same as TPR (true positive rate) $\frac{VP}{(VP+FN)}$.

F1 score
F1 score is a metric that tries to combine both Precision and Recall.

$$F1 - score = \frac{2 \times \text{ recall } \times \text{ precision}}{\text{recall+precision}}.$$

Modeling of churn in case of imbalanced class

The models are obtained with the best set of parameters. The hyperparameters have been adjusted by grid-search. The grid-search is an optimization method that will allow us to test a set of parameters and compare the performances to deduce the best parameterization.

Result based on the original database

Once the classifiers are trained, their performance is compared using the confusion matrix to determine accuracy, recall and f1-score.

TABLE 4 – Summary of results on the unbalanced base

Model	Accuracy	Precision	Recall	Score-F1	ROC AUC
Logistic regression	0.931	0.106	0.163	0.129	0.618
Decision tree	0.804	0.058	0.347	0.100	0.572
Random forest	0.699	0.056	0.541	0.101	0.635
Gradient Boosting	0.883	0.076	0.245	0.116	0.620
Support Vector Machine	0.877	0.077	0.265	0.119	0.639

These methods obtain an AUC value of about 0.62, which is low. According to the f1 score, we find that the predictive ability of our models is poor.

Result based on the resampling base

The methods that can be used to resolve unbalanced data are classified into two categories : the oversampling and undersampling approaches. undersampling is to undersample the majority class or oversampling is to oversample the minority class in the data set.

In this study, 70 % of the observations were used for learning, and the remaining 30 % were used for the test base. As a reminder, the learning base is made up of 242 terminated firms and 7,017 non-terminated companies. The proportion of positive cases is 3.3 %. The termination distribution after resampling is as follows :

TABLE 5 – Summary of results with under-sampling

	Non-Churn (= 0)	Churn (= 1)	Total
Imbalanced base	7 017 (96.7 %)	242 (3.3 %)	7 259
Oversampling	7 017 (50 %)	7 017 (50 %)	14 034
Undersampling	242 (50 %)	242 (50 %)	484
SMOTE	7 017 (50 %)	7 017 (50 %)	14 034
SMOTE &TomekLinks	6 995 (50 %)	6 995 (50 %)	13 990

The precision refers to the percentage of customers that the model expects to terminate and that have to terminate and those who have actually terminated. Therefore, we can say that the accuracy shows how certain we can be of positive predictions. It does not, however, say anything about the customers who will but whom the model incorrectly classifies as not terminated. The models produce a fairly low accuracy (see table above).

The recall, on the other hand, is interested in the customers who will actually terminate, and how many of them have been found. It is also an important measure for understanding how models can predict who will terminate.

Since it is more costly to not predict a resilient sample than to falsely predict a non-resilient sample, our interest is to try to correctly predict the resilient class. With the use of the resampling technique, we want to increase the number of true positives and decrease the number of false negatives, i.e. we want to increase recall.

TABLE 6 – Summary of results with under-sampling

	Recall				
	Imbalanced base	Oversampling	Undersampling	Smote	Smote & TomekLinks
Logistic regression	0.163	0.519	0.910	0.038	0.01
Decision tree	0.347	0.010	0.903	0.558	0.077
Random forest	0.541	0	0.923	0	0
Gradient Boosting	0.245	0.202	0.913	0.019	0.009
Support Vector Machine	0.265	0.250	0.817	0.250	0.26

Table 3 shows that using the subsampling technique increases the number of true positives (i.e., increases the number of correctly predicted resilient classes). However, for the unbalanced dataset, the random forest has the lowest score among the over-sampling ensemble methods, but showed higher scores for the original and undersampled datasets.

Modeling the elasticity of termination at the indexation rate

The elasticity of the probability of termination with respect to the indexation rate is the relative variation of this probability when the indexation rate varies by an infinitesimal amount. This allows us to know the reaction of the insureds to a variation in the indexation rate, in terms of cancellations.

Figure 4 shows us the distribution of elasticity from 2019 to 2021. In summary, the elasticity with respect to the indexation rate is generally positive and statistically significant, but its value ranges from 0 to 0.2 in 2019 and in 2020, and increased in 2022. These results indicate that people are paying attention to the change in indexation rate they face and their own risk of needing benefits.

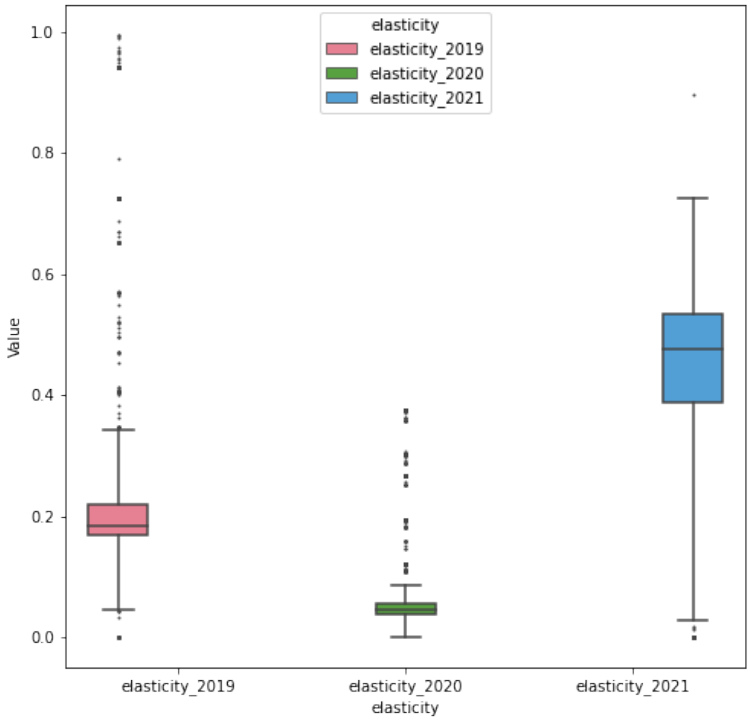

FIGURE 4 – Average elasticity of churn rate

Conclusion

In this dissertation, the task is to identify the customer profiles which have a higher probability of termination. After a comparison between the different learning algorithms, we obtained a conclusion : for our dataset, the results are not satisfactory as we expected. In order to improve the predictive performance of our models, we proposed a methodology on behalf of building up the prediction of the terminated (minority) class (the group of customers who chose to terminate their contract) by using resampling techniques. We proved that the predictions provided by the downsampling approach have the best performance. In the next part which studies on the elasticity, the logistic model helps to provide coefficients for measuring the elasticity of termination at the indexation rate. The study of elasticity allows us to optimize the tariff revision process, and also minimize the risk of termination.

Table des matières

Introduction

En assurance santé collective, avant la mise en place de la loi dite résiliation infra-annuelle (RIA), les assurés ne peuvent que résilier à leurs échéances annuelles, en adressant une lettre recommandée à l'organisme d'assurance au moins deux mois avant la date d'échéance. La loi dite résiliation infra-annuelle donnera la possibilité aux assurés de contrats collectifs de résilier à tout moment leur contrat de complémentaire santé, une fois la première année de souscription passée. À couverture équivalente, il semble donc raisonnable de supposer que les assurées soient tentées de mettre fin à leur(s) contrat(s), s'ils considèrent les prix sont trop élevées par rapport à ce que proposerait la concurrence. Les assureurs et mutuelles s'attendent désormais à une concurrence exacerbée. Pour appliquer une politique de rétention et pour garder l'équilibre de portefeuille. Il est important de mieux appréhender ce risque de résiliation.

Pour les assureurs, il y a deux éléments cruciaux pour réduire ce risque de résiliation, afin de mieux cerner son portefeuille client. Le premier est de pouvoir trouver des éléments explicatifs de ce comportement. Le second est capable de modéliser et prédire la résiliation d'un client sur un portefeuille donné.

Ce mémoire a deux objectifs, le premier est de prédire la résiliation d'un client sur un portefeuille donné et de pouvoir trouver des éléments explicatifs de ce comportement. Le deuxième est de modéliser l'élasticité-indexation individuelle pour répondre aux problématiques d'optimisation tarifaire.

Le mémoire s'articule en quatre parties :

1. la Partie I présente le contexte assurantiel de l'étude : l'assurance santé en France, ses contextes réglementaires, ainsi que l'apport du mémoire.
2. la Partie II traite notamment de l'aperçu des jeux de données sur lesquels les travaux sont effectués, ainsi que les analyses statistiques. Ces analyses effectuées sur le portefeuille de façon statique auront pour objectif de donner une première idée des variables qui influencent la résiliation.
3. la Partie III expose des rappels sur les principes des algorithmes d'apprentissage, les outils d'évaluation de la performance et les limites de ses structures prédictives. On termine cette partie par la présentation des différences entre les méthodes de ré-échantillonnages. Les détails sur les méthodes d'apprentissage automatique utilisées tout au long des différentes expérimentations constituent les points forts de cette partie.
4. la Partie IV présente le cas d'application, la prédiction de la résiliation et la mesure de l'élasticité au taux d'indexation. Dans un premier temps, nous chercherons à modéliser et à prédire la résiliation en fonction des variables explicatives retenues dans la base de données. Toutefois, les techniques de ré-échantillonage sont utilisées pour traiter le problème du déséquilibre des classes. Ensuite, nous construirons les premiers modèles de régression logistique pour mesurer l'élasticité de résiliation au taux d'indexation. Les différents résultats de nos expérimentations sont abordés. Le processus de comparaison de la performance des modèles, ainsi que la comparaison entre les métriques proposées. Enfin, une conclusion terminera le mémoire avec une ouverture sur des réflexions à explorer et les futurs travaux qui méritent d'être abordés.

Chapitre 1

Contexte de l'étude

Ce premier chapitre est consacré à la présentation de contexte de l'étude. Dans un premier temps, nous présentons le rôle de la Sécurité Sociale dans le système de protection sociale français. Le mécanisme de remboursement des frais de santé et l'assurance complémentaire santé collective. D'autre part, nous exposons le phénomène des résiliations et cadre général de la révision tarifaire chez MalakoffHumanis.

1.1 L'assurance Santé en France

L'assurance santé des français est composée de deux niveaux :

- Un régime de base ou régime obligatoire fourni par l'assurance maladie de la Sécurité sociale
- Un régime complémentaire proposé par les organismes assureurs [1].

La garantie « remboursement de frais de santé » (ou frais médicaux) a pour objet d'indemniser le salarié des dépenses occasionnées, pour lui-même, voire pour ses ayants droit (conjoints, enfants…), au titre de la santé. Il s'agit de compléter les prestations en nature servies par la sécurité sociale.

1.1.1 La notion de protection sociale

Créée en 1945, la Sécurité Sociale est l'acteur principal du système de santé français.

À l'origine, la Sécurité Sociale était réservée aux personnes salariées et leur famille. Elle s'est progressivement étendue à d'autres catégories de la population et a aujourd'hui pour principale mission de protéger les individus résidant en France des risques dits "sociaux" grâce à l'instauration de mesures collectives de protection et de prévention : c'est ce qu'on appelle la protection sociale.

Elle s'organise autour de trois logiques principales :

- la logique d'assurance sociale : couvrir une perte de ressource liée à un risque social ;
- la logique d'assistance : solidarité intergénérationnelle pour faire face aux diverses formes de pauvreté sociale ;
- la logique de protection universelle : 99,9 % de la population est couverte et les prestations sont accordées sans conditions de ressources ni de cotisations

1. Sociétés d'assurance Mutuelles ou Institution de prévoyance

Les différents régimes de la Sécurité Sociale
On a abouti ainsi à plusieurs « régimes » qui se distinguent entre eux principalement par le type de prestations qu'ils dispensent. On distingue en général quatre grands groupes de régimes sociaux :

- Le régime général : le Régime général concerne salariés et travailleurs assimilés à des salariés soit environ 80 % de la population.
  À noter : depuis le 1er janvier 2018, la protection sociale des travailleurs indépendants [2] est confiée au régime général de la Sécurité sociale.
- Le régime agricole : le régime agricole s'adresse à l'ensemble d'exploitants et salariés agricoles, ainsi que certains secteurs rattachés à l'agriculture (comme l'industrie agro-alimentaire).
- Le régime social des indépendants : Il couvre les artisans, commerçants, chefs d'entreprise et les professions libérales.
- Les régimes spéciaux : les régimes spéciaux regroupent des fonctionnaires, mineurs, militaires, agents de la SNCF, de l'EDF-GDF, étudiants, etc. En outre, depuis 1919, un régime spécial s'applique aux départements de la Moselle, du Bas-Rhin et du Haut-Rhin.

Les branches de la Sécurité Sociale
Une branche est une entité qui a à sa charge la gestion d'un ou plusieurs « risques ». La notion de risque peut être décrite comme le fait d'être exposé à un danger physique, comme la maladie, ou un danger économique, par exemple, le chômage. Le risque est alors dit « social » lorsqu'il entraîne une hausse des dépenses ou lorsqu'il provoque une baisse des ressources. La Sécurité sociale se compose de 5 branches :

- La branche maladie ;
- La branche accidents du travail et maladies professionnelles ;
- La branche famille ;
- La branche cotisation et recouvrement ;
- La branche retraite.

En particulier, la branche maladie assure la prise en charge des dépenses de santé des assurés et garantit l'accès aux soins. Elle favorise la prévention et contribue à la régulation du système de santé français. Elle recouvre les risques maladie, maternité, invalidité et décès.

Pour le régime général, la branche maladie est gérée par la Caisse nationale de l'Assurance Maladie et son réseau qui se compose des caisses primaires d'assurance maladie (CPAM), des caisses générales de sécurité sociale (CGSS) dans les départements d'outre-mer, des directions régionales du service médical (DRSM), des caisses d'assurance retraite et de la santé au travail (Carsat), ainsi que des unions de gestion des établissements de caisse d'assurance maladie (Ugecam).

Le financement de la protection sociale
La protection sociale était à l'origine financée exclusivement par des cotisations. Le financement s'est diversifié dans les années 1990 avec la CSG et d'autres impôts. Aujourd'hui, les ressources de la Sécurité sociale se répartissent en trois catégories :

2. (artisans, commerçants et professions libérales)

- Les cotisations sociales (env. 58 % des recettes) : assises sur le travail salarié (part salariale, part employeur) et sur les revenus de toute nature. Ce sont des ressources en diminution ces dernières années, mais elles restent la première source de financement ;
- La CSG (env. 20 %) : prélèvement opéré sur l'ensemble des revenus ;
- Autres impôts et taxes (env. 13 % des recettes) : ce sont les divers prélèvements de nature fiscale, contributions et taxes affectées au financement de la Sécurité Social (TVA sur les tabacs, taxes sur les salaires...) ;
- Autres sources : transferts en provenance de l'Etat, autres organismes (fonds solidarité vieillesse), autres régimes.

1.1.2 Mécanismes de remboursement de la Sécurité Sociale

L'assurance maladie intervient sur la base de tarifs fixés par convention ou d'autorité. Tout dépassement par rapport à ces tarifs est à la charge de l'assuré ou d'une assurance complémentaire.

Une participation dite « ticket modérateur » est laissée à la charge de l'assuré. Elle peut être proportionnelle ou forfaitaire, elle varie selon les catégories de prestations. Le schéma ci-dessous décrit la répartition des frais à la charge de l'assuré et ceux pris en charge par le régime de la Sécurité Sociale.

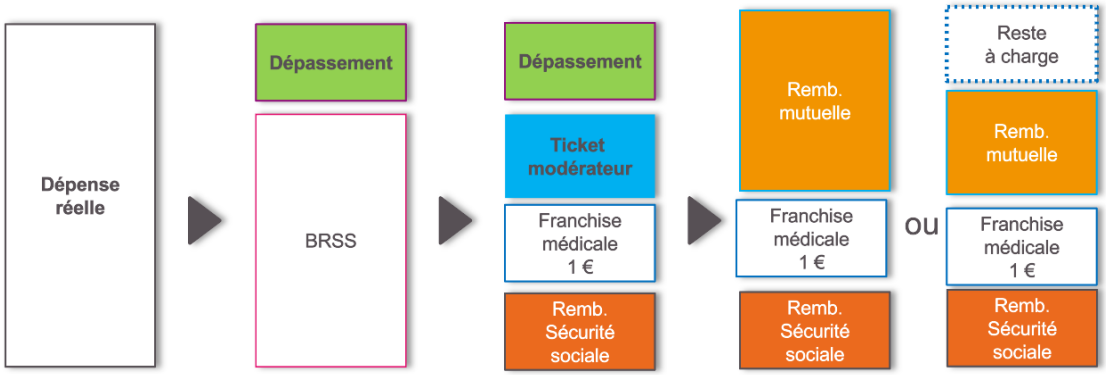

FIGURE 1.1 – Mécanisme remboursement Sécurité Sociale

Dépense réelle
Somme engagée par une personne pour ses dépenses de santé

BRSS
Tarif servant de base à la Sécurité sociale pour effectuer le remboursement des honoraires et des soins dispensés par les praticiens

Dépassement d'honoraires
Part du prix d'une consultation ou d'un acte médical qui dépasse la base de remboursement

RSS : remboursement Sécurité sociale
Montant remboursé par la Sécurité sociale

RM : remboursement Mutuelle
Montant remboursé par la mutuelle

Le reste à charge
La somme que l'assuré doit verser lorsque celui-ci a reçu les remboursements de la Sécurité Sociale et de sa complémentaire santé dite « Mutuelle »

TM : Ticket modérateur
Le ticket modérateur est la partie des dépenses de santé qui reste à charge par l'assuré une fois que l'assurance Maladie a remboursé sa part

1.1.3 Les grilles de garanties

Lorsqu'une personne souscrit à un contrat d'assurance santé, l'une des premières choses à analyser est la grille de garanties qui lui indique toutes les prestations dont il aura droit de la part de son organisme d'assurance.

Cette grille se présente sous la forme d'une liste organisée de prestations avec leur remboursement associé et ces prestations sont pour la plupart du temps classées selon des grands postes de soin. Une grille se présente donc souvent sous la forme de tableau en blocs. Les organismes sont libres sur la mise en forme et sur la segmentation de leurs grilles tout comme sur l'expression des garanties mais une uniformisation des grilles tend à se faire de plus en plus (contrat responsable, réforme 100 % santé, ...). En effet, avec la réforme 100 % santé par exemple, l'UNOCAM [3] et ses fédérations s'engagent à améliorer et à faciliter la lisibilité des garanties santé par le biais d'intitulés communs et harmonisés sur les principaux postes de prestations à partir du 1er janvier 2020. Il s'agit cependant seulement de recommandations et non d'obligations mais il est quand même observé que les mutuelles sont fortement encouragées à les respecter.

Les grands postes les plus courants sont :

- Les soins courants aussi appelés « frais médicaux » ou « soins de ville » constituent l'un des postes de dépenses les plus importants. Il comprend notamment les consultations chez le médecin ou le spécialiste ou encore les soins effectués par les auxiliaires médicaux (kinésithérapie, soins infirmiers, ...) ;
- L'hospitalisation est un poste principalement constitué des prestations liées à une visite ou opération à l'hôpital : frais de séjour, forfaits journaliers, chambre particulière... ;
- L'optique avec les montures, les verres, les lentilles, la chirurgie des yeux... ;
- Le dentaire avec les prothèses dentaires, l'orthodontie et les soins dentaires ;
- L'aide auditive avec les audioprothèses et autres appareillages auditifs.

1.1.4 Les différents acteurs

Le recours à des organismes complémentaires est indispensable pour subvenir aux frais de santé engagés en cas de maladie. Les organismes qui supportent le risque lié aux dépenses de santé sont listés ci-dessous :
Compagnie d'assurance :
Elle est régie par le Code des Assurances, elle intervient sur l'ensemble des domaines de l'assurance (Multi Risque Habitation, assurances de biens, santé, prévoyance et retraite).

Mutuelle :
Elle est régie par le Code de la Mutualité, ce sont des groupements à but non lucratif qui gèrent essentiellement des garanties Frais de santé. Elle repose sur le principe de solidarité entre l'ensemble des adhérents, qui sont également nommés "les sociétaires".

Institution de Prévoyance :
Elle est régie par le Livre IX du Code de la Sécurité Sociale, personnes morales de droit privé à but non lucratif dont l'activité principale est la prévoyance collective (santé et prévoyance).

3. (l'Union Nationale des Organismes Complémentaires d'assurance Maladie qui rassemble les différentes familles de complémentaires santé)

Ils sont particulièrement représentés en prévoyance du fait des clauses de désignations, jugées anticonstitutionnelles par le Conseil Constitutionnel depuis le mois de juin 2013.

Le graphique ci-dessous représente la proportion de contrats collectifs et contrats individuels, en 2020, en fonction des différents acteurs. Nous pouvons constater que la majorité des contrats santé souscrits au sein d'institutions de prévoyance porte sur les contrats santé collectifs, lesquels représentent 87 % des cotisations qu'elles ont collectées en 2020. Les mutuelles sont quant à elles largement positionnées sur les contrats santé individuels (68 % de leur activité). Les sociétés d'assurances sont dans une position intermédiaire, avec 54 % des cotisations collectées au titre de contrats collectifs.

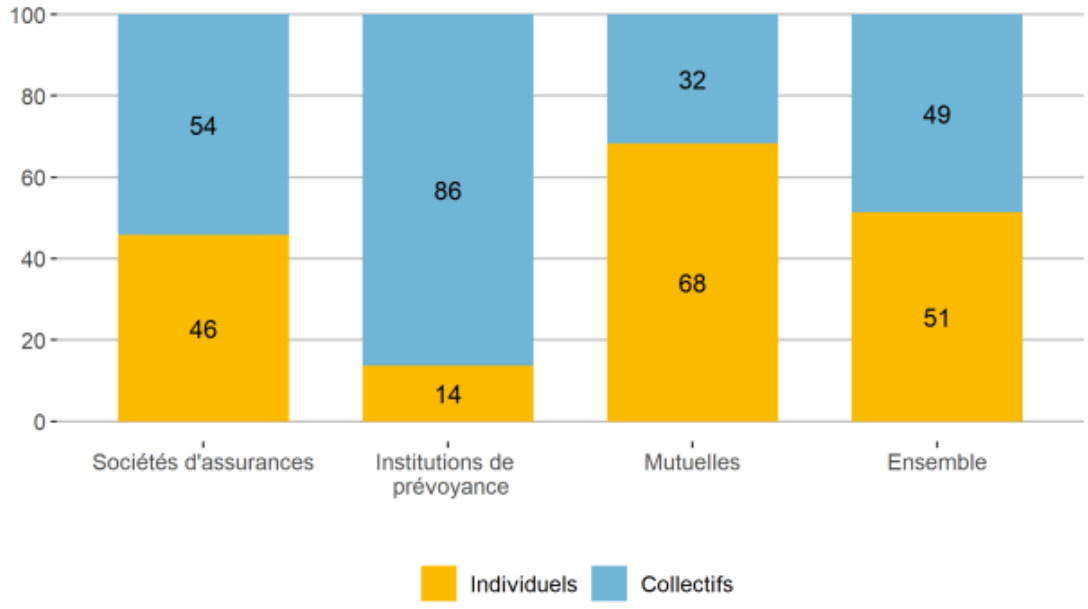

FIGURE 1.2 – Part des contrats collectifs et individuels dans l'ensemble des cotisations collectées en santé par les différents types d'organismes en 2020

1.1.5 La complémentaire santé

En France, on peut noter que dès sa création en 1945, l'Assurance maladie obligatoire de la Sécurité sociale imposait aux assurés sociaux un reste à charge prenant la forme d'une fraction du tarif conventionnel. Depuis 1945, l'assurance complémentaire santé permet de compléter les remboursements des actes de santé de l'Assurance Maladie obligatoire, en totalité ou en partie, afin de réduire ainsi les restes à charge auxquels sont soumis les individus.
En effet, l'Assurance maladie ne rembourse qu'une partie des dépenses santé. Certains soins, comme la médecine douce ou l'orthodontie pour adultes, ne sont pas du tout pris en charge.
La mutuelle ou l'assurance santé d'entreprise est un contrat collectif qui sert à protéger l'ensemble des salariés d'une entreprise.
Cette complémentaire santé fixe différents niveaux de garanties en fonction des actes de santé. Elle est adaptée au besoin moyen des salariés de l'entreprise et choisie par l'employeur. Ainsi, celle-ci peut ne pas convenir à un salarié ou répondre à ses besoins propres.

Un contrat d'assurance maladie complémentaire, souvent abrégé par abus de langage en « complémentaire santé » peut être collectif ou individuel. Un contrat collectif couvre un groupe de salariés dans une entreprise

et si le contrat le permet, leurs ayants droit. C'est un contrat souscrit par l'employeur pour l'ensemble de ses salariés ou pour une catégorie objective de salariés (ex : les cadres). Depuis le 1er janvier 2016 avec l'entrée en vigueur de la généralisation de la complémentaire santé, il y a obligation de couverture de ses salariés par l'entreprise. Un contrat individuel couvre un particulier et parfois aussi ses ayants droit. Il est à souscription libre et à accès personnel.

La figure 1.3 ci-dessous indique le fonctionnement de 3 acteurs et le contrat collectif et le contrat individuel.

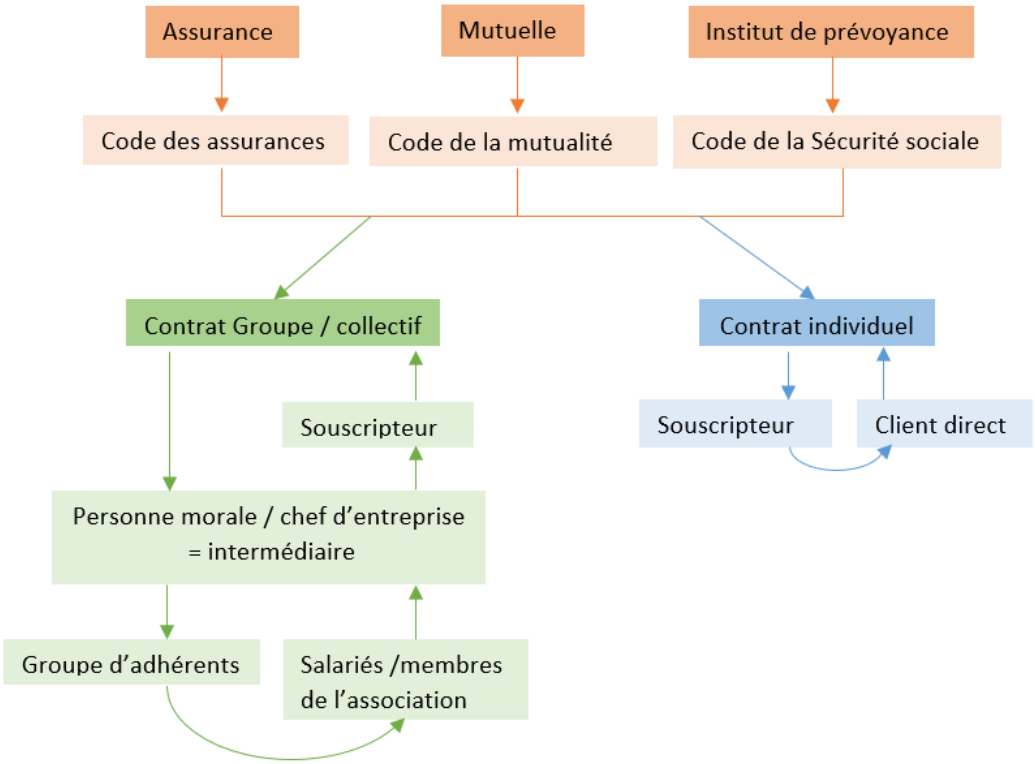

FIGURE 1.3 – schéma-contrat-complémentaire-santé

Pour un contrat collectif, on distingue les contrats facultatifs et les contrats obligatoires :

- Collectif facultatif : l'adhésion est individuelle et facultative de la part des bénéficiaires du contrat ;
- Collectif obligatoire : mis en place dans le cadre de la généralisation de la complémentaire santé en entreprise, le contrat santé collectif et obligatoire concerne toutes les entreprises du secteur privé.

Dans le cadre du présent mémoire, c'est le contrat collectif obligatoire mis en place dans les entreprises que nous allons analyser plus précisément.

1.1.6 Contrat responsable

Les contrats dits « responsables et solidaires » sont entrés en vigueur en 2006. Les conditions ont évolué en 2015 et avec l'arrivée du 100 % santé en 2019, mais l'objectif des contrats de complémentaire santé qui répondent à ces exigences reste le même :

- inciter les patients à respecter le parcours de soins coordonnés afin de bénéficier d'une bonne prise en charge de leurs dépenses de santé ;

- limiter les dépenses de santé publique.

les avantages des contrats responsables
Le régime fiscal et social des contrats responsables est plus favorable, puisqu'il permet de bénéficier :

- d'une exonération des charges sociales pour les cotisations versées par l'employeur dans la limite d'un certain plafond si le contrat est collectif et obligatoire,
- d'une taxe de 13,27 % (au lieu de 20,27 % pour les contrats non responsables),
- pour les travailleurs non-salariés : d'une fiscalité avantageuse (déductibilité de la cotisation pour l'impôt sur le revenu),
- pour les travailleurs salariés : d'une déduction de la part salariale dans le calcul de l'impôt sur le revenu.

1.1.7 Les dispositifs de solidarité

Il existe des alternatives qui sont possibles grâce au principe de solidarité afin de donner l'accès aux soins aux individus ne disposant pas de complémentaire santé et/ou ceux n'ayant ni le statut, ni une activité professionnelle leur permettant de cotiser et d'accéder à une couverture sociale normale :

- La Protection Universelle Maladie (PUMA) : entrée en vigueur le 1er janvier 2016, elle permet à toute personne qui travaille ou réside en France de manière stable et régulière, un droit à la prise en charge de ses frais de santé à titre personnel et de manière continue tout au long de la vie ;
- Aide Médicale d'État (AME) : elle permet aux ressortissants étrangers en situation irrégulière sous conditions de résidence stable et de ressources la prise en charge de certains soins et actes ;
- La Couverture Maladie Universelle Complémentaire (CMU-C) qui est une sorte de complémentaire santé pour les personnes ayant un revenu modeste ;
- L'aide à l'acquisition d'une Complémentaire Santé (ACS) qui est une aide financière attribuée par le régime général de santé lorsque le bénéficiaire dispose de revenus faibles mais pas suffisamment pour obtenir la CMU-C. La CMU-C et l'ACS ont cependant fusionné le 1er novembre 2019 pour devenir la Complémentaire Santé Solidaire (CSS) et bien que gratuite pour les bénéficiaires de la CMU-C, elle demande une participation financière pour les anciens bénéficiaires de l'ACS. Les garanties de la CMU-C sont gardées.

1.1.8 Contrat collectif obligatoire

En effet dans ce type de contrat une personne morale ou un chef d'entreprise va souscrire un contrat auprès d'une société d'assurance, d'une mutuelle ou d'un institut de prévoyance. La relation contractuelle relative aux garanties présentes dans le contrat et à son fonctionnement est donc signée entre ce souscripteur et l'assureur. La troisième partie est donc l'assuré. Il est lié directement à l'intermédiaire ayant signé le contrat que ce soit sur la base d'une relation employeur/employé dans le cadre de la mutuelle d'entreprise obligatoire.

Pour la mise en place du contrat collectif, la loi du 8 août 1994 précise les 3 seules modalités envisageables pour la mise en place de garanties collectives :

- Conventions ou accords collectifs ;
- Référendum ;
- Décision unilatérale du chef d'entreprise dite « DUE ».

L'accord collectif est comparable dans ses modalités et effets, à ce qu'il est pour d'autres aspects de la relation de travail :

- négocié et signé avec les délégués syndicaux (le Comité d'entreprise n'a aucune légitimité en matière de négociation de ce type) ;

- s'impose à tous les salariés ;
- ne peut être modifié que par un texte de même type.

L'accord référendaire est soumis au vote de tous les salariés :

- nécessite le respect d'un certain formalisme (type élections du personnel avec information préalable, bureau de vote, dépouillement organisé) ;
- s'impose à tous les salariés ;
- ne peut être modifié que par un texte de même type ou un accord collectif.

La décision unilatérale de l'employeur est très souvent utilisée dans les petites entreprises, mais attention :

- nécessite également un certain formalisme pour être « conforme » : elle doit être matérialisée par un écrit remis à chaque salarié ;
- ne peut en aucun cas imposer un précompte salarial de cotisation, chaque salarié présent dans l'entreprise lors de la mise en place du régime restant libre de refuser celui-ci lors de la mise en place du régime (article n°11 de la Loi EVIN).

Le contrat souscrit par l'employeur doit respecter les 4 critères essentiels suivants :

- collectif : obligation de couvrir tous les salariés d'une même branche professionnelle, d'une même entreprise ou d'une même catégorie objective de salariés et ce avec le même niveau de couverture ;
- obligatoire : obligation pour l'ensemble des salariés d'adhérer au contrat lorsqu'il est mis en place au sein d'une entreprise ;
- responsable : obligation de respecter un certain nombre d'interdictions de prise en charge de frais de santé ;
- solidaire : interdiction de fixer les cotisations en fonction de l'état de santé des individus couverts et de recueillir des informations médicales à l'adhésion du contrat.

1.2 Contexte réglementaire

1.2.1 Solvabilité II

Comme toutes les entreprises d'assurance sont soumises à des exigences de contrôle et de réserves en capital, pour faire face aux risques assurés (exigences dites de « solvabilité »). Les exigences en capital pour couvrir les risques ont des racines très anciennes datant de la fin du 19ème siècle ([Jaouen, 2019], [Lambert, 1998]).

Les fusions progressives des activités étatiques de contrôle ont par ailleurs conduit aujourd'hui à une autorité unique pour les banques et les assurances : l'Autorité de Contrôle Prudentielle et de Résolution (ACPR). Le contrôle est dit « prudentiel » et l'organisme est aussi en charge de mesures de « résolution » : l'ACPR « doit également veiller à l'élaboration et la mise en œuvre des mesures de prévention et de résolution des crises bancaires » [Jaouen, 2019].

Les règles sont harmonisées au niveau européen. Elles visent tout autant la protection de la clientèle que la stabilité du système monétaire et financier ainsi que la prévention des crises. Mais cette harmonisation a aussi pour objectif de faciliter et développer le fonctionnement du marché intérieur dans le domaine des assurances [Jaouen, 2019].

1.2.2 La loi Évin

La loi Évin a marqué une étape fondamentale dans la réglementation de la protection sociale complémentaire.

En cas de rupture du contrat de travail, l'article 4 de la loi Évin du 31 décembre 1989 n° 89-1009 prévoit la possibilité de maintenir les garanties collectives frais de santé, sous certaines conditions.

Dans ce cadre, les garanties santé sont maintenues sans condition de durée. Il suffit que l'ancien salarié en fasse la demande dans les 6 mois qui suivent la rupture du contrat de travail ou la fin de ses droits à portabilité et qu'il verse les cotisations nécessaires.

Selon l'article 4 de la loi Évin sur la mutuelle obligatoire collective, les anciens salariés peuvent continuer à bénéficier de cette mutuelle lorsqu'ils ont quitté l'entreprise. Les personnes concernées sont :

- les anciens salariés bénéficiaires d'une rente d'incapacité ou d'invalidité ;
- les bénéficiaires d'une pension de retraite ;
- les personnes qui perçoivent les allocations chômage ;
- les ayants droit d'un salarié décédé qui veulent conserver la mutuelle collective du défunt.

Depuis le 1er juillet 2017, l'encadrement de la cotisation est un peu différent pour les anciens salariés qui demandent à bénéficier du maintien de leurs garanties santé :

- La première année, les tarifs ne peuvent être supérieurs aux tarifs globaux applicables aux salariés actifs ;
- La deuxième année, les tarifs ne peuvent être supérieurs de plus de 25 % aux tarifs globaux applicables aux salariés actifs ;
- La troisième année, les tarifs ne peuvent être supérieurs de plus de 50 % aux tarifs globaux applicables aux salariés actifs.

1.2.3 L'ANI

La loi ANI (Accord National Interprofessionnel) est entrée en vigueur le 1er janvier 2016, cet accord indique qu'une couverture frais de santé collective et obligatoire minimale doit être proposée à tous les salariés par les entreprises. Le contrat mis en place doit respecter les critères du contrat responsable pour ouvrir droit aux bénéfices des avantages sociaux et fiscaux.

Cette couverture collective obligatoire doit remplir les conditions suivantes :

- La participation financière de l'employeur doit être au moins égale à 50 % de la cotisation. Le reste est à la charge du salarié ;
- Le contrat doit respecter un socle de garanties minimales, aussi appelé panier de soins minimum ;
- Le contrat est obligatoire pour les salariés, sauf dans les cas où le salarié peut refuser la mutuelle.

La portabilité de la complémentaire santé
En cas de cessation du contrat de travail non consécutive à une faute lourde, le salarié couvert par un contrat collectif a droit à son maintien à titre gratuit pendant une durée égale à la période d'indemnisation du chômage, dans la limite de 12 mois (article L. 911-8 du code de la sécurité sociale, issu de la loi du 14 juin 2013 relative à la sécurisation de l'emploi). Le financement de cette gratuité est entièrement mutualisé, c'est-à-dire qu'il est assuré par les cotisations perçues au titre des salariés restant dans l'entreprise (article 2 de l'ANI du 11 janvier 2013).

Pour bénéficier de l'article 4 de la loi Evin sur la mutuelle et la portabilité, les anciens salariés disposent de 6 mois à compter de la rupture du contrat pour faire leur demande. Les ayants droit disposent également de 6 mois après le décès pour demander la portabilité de la mutuelle. Les anciens salariés qui demandent la portabilité obtiennent un maintien des garanties sans condition de durée. En revanche, les ayants droit ne peuvent bénéficier de la loi Evin sur la mutuelle que durant 12 mois. Notez que continuer à bénéficier des garanties de la mutuelle collective est facultatif. L'ancien salarié peut parfaitement renoncer à la portabilité et choisir sa propre mutuelle individuelle.

Différences entre la loi Evin et la loi ANI

Il existe plusieurs différences entre la loi Evin sur la mutuelle et la loi ANI. La loi Evin, hormis pour les ayants droit d'un salarié décédé, permet un maintien des garanties sans condition de durée. Avec la loi ANI, la portabilité ne peut excéder 12 mois et elle est calculée en fonction de l'ancienneté du salarié. S'il était dans l'entreprise depuis plus d'un an, il pourra bénéficier du maintien des garanties durant 12 mois. S'il n'est resté que 6 mois, la portabilité ne durera alors que 6 mois également. Deuxième point important : les cotisations. La loi Evin sur la mutuelle des retraités, personnes en incapacité ou invalidité, et demandeurs d'emploi, prévoit que l'ancien salarié paiera lui-même ses cotisations. Dans le cadre de la loi ANI, l'ancien salarié, tout au long de la période de portabilité, bénéficiera à la fois du maintien des garanties et de la part de cotisation de l'entreprise.

	Loi Evin	ANI
Couverture	Santé	Santé et prévoyance
Assuré(s)	Le salarié	Le salarié et ses ayants droit
Durée	Dispositif viager	Durée temporaire de 12 mois maximum
Point de départ	Le lendemain de la date de la demande du salarié	Date de la cession du contrat de travail

TABLE 1.1 – Loi Evin et Accord National Interprofessionnel (ANI)

1.2.4 Le 100 % santé

Lors de sa campagne présidentielle de 2017, Emmanuel Macron avait promis une réforme du système de santé français avec l'ajout du « reste à charge zéro » (connu à présent sous le nom de « 100 % Santé » pour des raisons de communication). Promesse tenue puisque fin octobre 2018, l'article 33 de la Loi de Financement de la Sécurité Sociale (LFSS) 2019 est adopté par l'Assemblée générale et la mise en œuvre progressive du 100 % Santé dès 2019 est mise en marche.

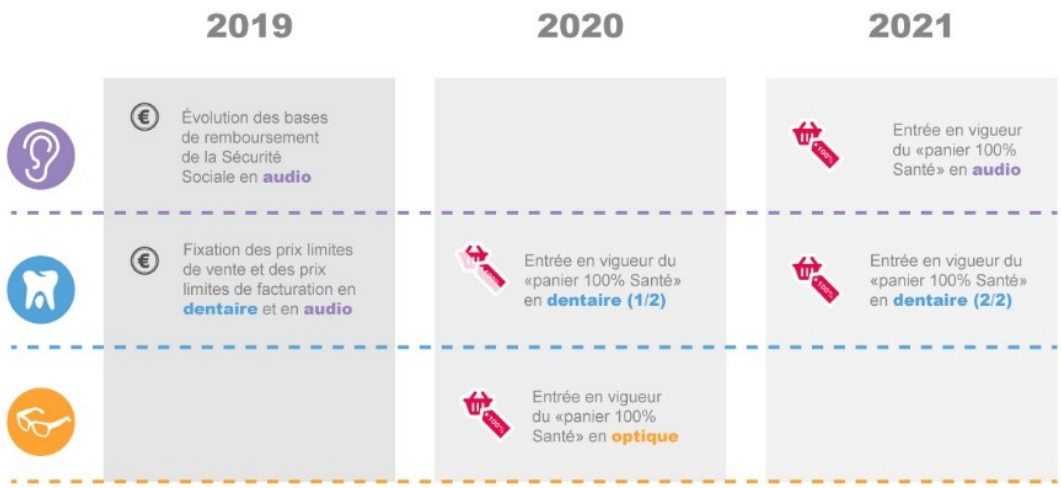

FIGURE 1.4 – 100 % santé illustration

Le gouvernement a mis en place une politique visant à obtenir un reste à charge nul sur certains postes de soins (optique, dentaire, audioprothèse) pour éviter le renoncement aux soins. Elle a nécessité de définir des paniers de soins « 100 % santé » dans les trois domaines avec l'objectif d'une bonne qualité et de prix maîtrisés.

Pour réaliser cet objectif, la réforme s'appuie sur trois leviers : une modification des bases de remboursement de la Sécurité Sociale (tendance haussière pour une majorité des équipements), des prix limites de vente des équipements et le reste à charge nul.

Le gouvernement a souhaité obtenir par la négociation la mise en place d'une logique de cofinancement entre l'assurance maladie et les complémentaires santé, tout en exigeant de ces dernières que les tarifs n'augmentent pas.

Cette réforme a impacté le monde de la santé dans son intégralité : les assurés qui se voient ouvrir une nouvelle alternative de soins sans frais, la Sécurité Sociale et les assureurs qui doivent prendre à leur charge plus que ce qu'ils ont l'habitude de payer et les professionnels de santé qui se voient obligés de proposer ces paniers « zéros » dans leur devis avec des prix limites de vente pratiqués et de changer leur manière de communiquer.

1.2.5 Résiliation infra-annuelle

Depuis le 1er janvier 2016, les entreprises du secteur privé ont l'obligation de proposer à leurs salariés un régime complémentaire santé collectif bénéficiant d'un dispositif social et fiscal de faveur selon des règles liées à la formalisation du régime lui –même ainsi qu'à l'encadrement des garanties proposées.
Ce type de contrat, souscrit en général à effet du 1er janvier pour une durée de 12 mois, est renouvelable annuellement par tacite reconduction sauf dénonciation à formaliser dans un délai de 2 mois précédant la date de renouvellement.

Tacitement reconduit chaque année, il est désormais possible de résilier à tout moment un contrat de complémentaire santé, sans frais ni pénalité, dès lors que le contrat a été souscrit depuis au moins un an. Cette possibilité est entrée en vigueur au 1er décembre 2020, instituée par la loi du 14 juillet 2019 relative au droit

de résiliation sans frais de contrats de complémentaire santé et le décret n°2020-1438 du 24 novembre 2020, aussi appelée résiliation infra annuelle.

La faculté de résiliation infra-annuelle est applicable aux contrats santé, en cours ou nouvellement souscrits, auprès d'un organisme d'assurance, d'une mutuelle ou d'une institution de prévoyance. L'ensemble des contrats de complémentaire santé sont concernés qu'ils soient responsables ou non, individuels ou collectifs, à adhésion obligatoire ou facultative.

Condition de délai pour pouvoir résilier à tout moment son contrat de complémentaire santé
Le souscripteur ne peut résilier son contrat qu'après l'expiration d'un délai d'un an à compter de la 1ère souscription. La résiliation prend alors effet un mois après que l'assureur en a reçu notification.

Par exemple, si le contrat est souscrit le 01/01/2021, la résiliation ne peut être adressée qu'à compter du 01/01/2022. La résiliation prendra effet un mois après la date de réception de la notification. Cette date est présumée être le premier jour qui suit la date d'envoi de cette notification qui figure sur le cachet de La Poste de la lettre recommandée. Dans le cas, la résiliation prendrait effet le 2 février 2022.

La loi s'applique aux contrats en cours au 1er décembre 2020. Ainsi, tous les contrats entrant dans le périmètre précisé question 1, même conclus avant le 1er décembre 2020, peuvent être résiliés à tout moment.

FIGURE 1.5 – Résiliation infra annuelle Illustration

1.3 Contexte général de révision tarifaire

Révision annuelle des tarifs des contrats existants par Malakoff Humanis, c'est une étape qui intervient le plus souvent une fois par an, à la date d'échéance anniversaire du contrat, lors d'une éventuelle revalorisation des primes. C'est donc au cours de cette étape, nous nous intéresserons à l'impact du tarif sur le comportement de l'assuré en déterminant sa propension à résilier en fonction de son tarif.

Deux principaux mécanismes de révision tarifaire :

- L'indexation (ou RT1) : une revalorisation uniforme de l'ensemble des contrats afin de refléter l'inflation, les évolutions réglementaires
- Le redressement (ou RT2) : révision des tarifs ou des garanties pour les contrats présentant un déséquilibre trop important entre les prestations versées et les cotisations perçues

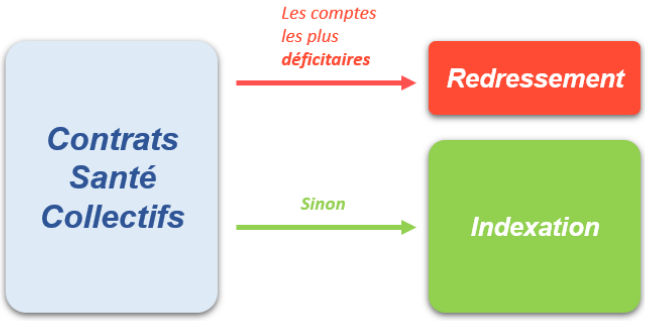

FIGURE 1.6 – Mécanisme remboursement Sécurité Sociale

les taux redressés sont calculés en fonction des objectifs du groupe et révisés par des différents interlocuteurs. En revanche l'indexation est un processus systématique et les taux sont homogènes. Donc dans notre étude, on ne s'intéresse qu'à l'indexation, mais pas le redressement.

1.3.1 Modélisation de la résiliation en cas de déséquilibre des classes

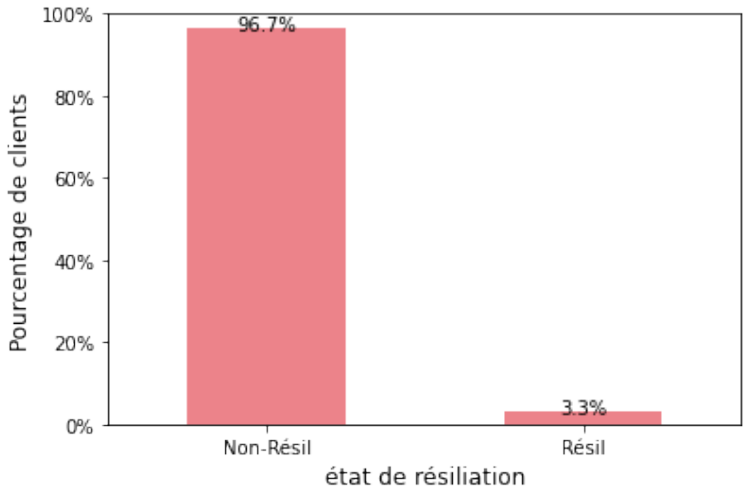

FIGURE 1.7 – La répartition de nombre de clients en fonction de la variable cible résiliation en 2021

Le premier modèle de notre étude est d'anticiper les résiliations des entreprises, dans notre cas, une très grande majorité des entreprises ne sont pas résiliées. Les données déséquilibrées ont causé des difficultés pour développer un bon modèle de prédiction. Si ce facteur n'est pas pris en compte, on risque de construire un modèle qui sera incapable de détecter les résiliations. Cette étude a appliqué une combinaison de techniques d'échantillonnage pour améliorer le modèle de prédiction des résiliations des clients.

1.3.2 Mesure de l'élasticité de résiliation au taux d'indexation

Afin de lier la révision tarifaire et le phénomène de résiliation, nous essaierons d'étudier l'impact d'une augmentation du taux d'indexation sur l'attrition des assurés. Cet impact est connu sous le nom d'élasticité de la probabilité de résiliation au taux d'indexation.

Chapitre 2

La présentation des données

Pour réaliser nos modèles de résiliation, l'étape préliminaire indispensable est de construire une base de données robuste la plus complète possible contenant les informations des contrats santé standard collectifs, sur plusieurs années, afin d'obtenir un maximum de caractéristiques concernant le contrat, le taux d'indexation notamment. Nous préférons dans un premier temps sélectionner un large nombre de variables potentiellement explicatives à notre disposition, quitte à réaliser une sélection plus fine par la suite, après analyse des statistiques descriptives.

2.1 Description des données brutes

Les données utilisées dans le cadre de ce mémoire sont issues d'un portefeuille de contrats de complémentaire santé collectif standard [1].

Les données relatives aux affiliations

La base contrat contient des informations sur tous les contrats ayant été actifs entre 2014 et 2022. Par ailleurs, elles permettent de récupérer les informations concernant les dates de prise d'effet du contrat et les dates de résiliation lorsque le contrat a été résilié.

- Numéro de contrat : le numéro permettant d'identifier le contrat. Il reste le même durant toute la durée de vie du contrat,
- Type de produit souscrit,
- Les dates d'effet et de résiliation éventuelle des contrats,
- Le segment : il s'agit du type de client avec deux possibilités, Standard ou Sur-mesure, dans ce mémoire, on se concentre sur le segment Standard.

Les données relatives aux cotisations

La table de cotisation contient l'ensemble des cotisations encaissées de 2014 à 2022 qui correspondent au montant hors taxes que l'assuré a versé.

Les données relatives aux prestations

Les variables permettant d'identifier le sinistre :

- Numéro de contrat : le numéro permettant d'identifier le contrat. Il reste le même durant toute la durée de vie du contrat,
- Date de survenance : c'est la date à laquelle la compagnie a enregistré le sinistre,

1. les offres standards sont de manière générale dédiées au TPE (1 à 9 salariés) et aux petites PME (10 à 49 salariés)

- Garantie sinistrée : hospitalisation, optique,
- Coût du sinistre : il s'agit du montant payé par la compagnie dans le cadre du sinistre survenu.

Les données relatives aux taux d'indexation
La cotisation d'un contrat de complémentaire santé évolue chaque année, lors de la mise en place des tarifs, il est possible, contractuellement, d'exprimer les cotisations en pourcentage d'indices. Donc la table de taux d'indexation est obtenue à partir de la table tarif. Ce taux est défini en fonction des garanties et des différents assiettes, les trois assiettes sont : forfaits, pourcentage du salaire, et le plus couramment est pourcentage du PMSS (Plafond Mensuel de la Sécurité Sociale).

Pour mieux comprendre avec un exemple, imaginons l'évolution du PMSS est de 1 % entre deux années. Le taux d'indexation de son contrat d'assurance santé est de 4 %. L'indexation appliquée future sera alors de 5.04 %. [2]

Les données résiliations
La table des résiliations permettant d'obtenir l'information de la résiliation de l'assuré, mais aussi les motifs. Dans la base résiliation, nous avons deux motifs de résiliation qui viennent de deux sources différentes : gestion et commercial.

Les variables qui nous intéressent ici sont les suivantes :

- L'année de l'observation
- Le numéro d'entreprise
- La date de résiliation : date à laquelle le contrat a été résilié.
- La date de réception de courrier de résiliation
- Le motif de résiliation (Commercial)
- Le motif de résiliation (Gestion)
- L'initiative de la résiliation (par le client ou la compagnie)

En effet, il faut discerner la date de réception de courrier de résiliation, de la date d'effet de la résiliation qui indique le jour où le contrat prend réellement fin. Ces deux dates de résiliation sont intéressantes pour étudier le comportement des clients en observant le temps qui s'écoule entre ces deux dates.

2.2 Construction de la base de modélisation

Afin de construire une base de modélisation fiable et exploitable, de nombreux traitements ont été réalisés sur les données brutes décrites précédemment. Ils seront détaillés dans cette section.

Les bases de données présentées auparavant ont été jointes grâce aux colonnes en commun : l'année de l'observation, l'identifiant de l'entreprise, le numéro de contrat, ce sont des colonnes appartenant à la clé primaire de chacune des tables. Une clé primaire permet d'identifier de manière unique un enregistrement dans une table. Elle permet notamment d'effectuer des opérations de jointures entre les tables.

Les traitements intermédiaires aux jointures reposent sur l'objectif fondamental suivant : la base finale doit comporter une ligne par entreprise par an.
Ils peuvent être synthétisés par le schéma ci-contre :

2. $(1+0.04)(1+0.01)-1 = 0.0504$

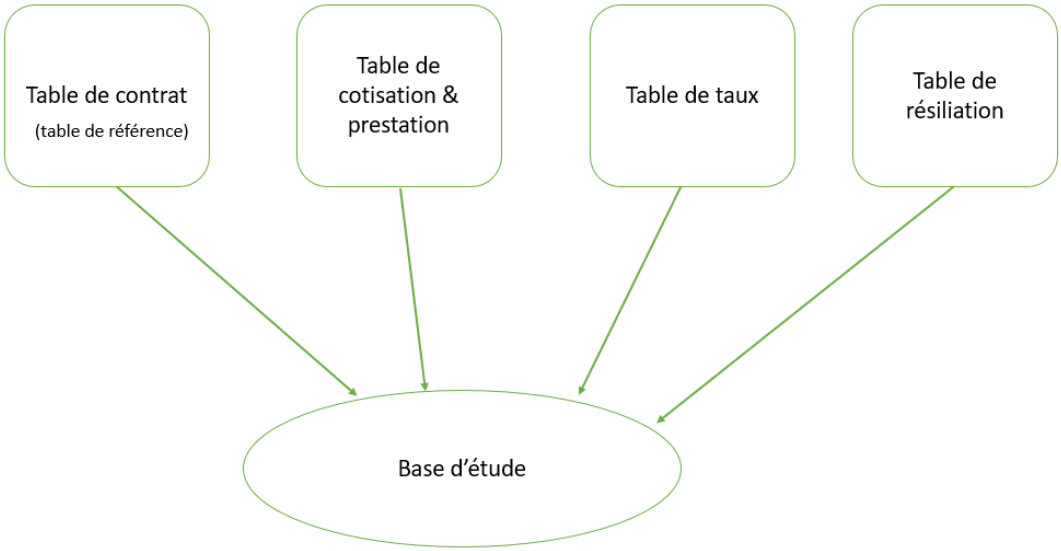

FIGURE 2.1 – Regroupement des bases

Les lags
Les lags permettent de prendre en compte les historiques de l'entreprise. La création de lags de variables est une méthode couramment utilisée pour la prédiction de phénomènes temporels. Il s'agit de créer de nouvelles variables correspondant aux variables d'origine mais décalées d'un certain nombre de lignes par l'écart de temps souhaité. Il permet de voir l'évolution des cotisations et prestations etc au fur à mesure des année. Ce principe est illustré dans l'exemple ci-dessous :

Année	Siren	Cotisation_nm0	Cotisation_nm1	Cotisation_nm2	Cotisation_nm3
2019	12 345	140	99	88	110
2020	12 345	150	140	99	88
2021	12 345	200	150	140	99
2022	12 345	100	200	150	140

TABLE 2.1 – Exemple de lag

2.3 Périmètre d'étude

Afin de travailler sur une base de modélisation fiable et cohérente, il est important de bien définir les paramètres d'étude. D'abord, il n'est pas possible d'utiliser les informations de 2022 car qu'elles ne sont pas complets. Donc on s'intéresse aux historiques de 2021, 2020, et 2019.

Pour la modélisation de la résiliation, nous allons étudier les résiliations de 2021, et nous cherchons à construire des modèles pour prédire la résiliation en 2021, sur le périmètre santé collectif standard.

2.3.1 Choix de la période d'observation

Du fait que chaque année, chez Malakoff Humanis, l'assureur informe sa clientèle par courrier du nouveau tarif de l'année NM0 le 30 octobre NM1. L'assuré peut donc prendre sa décision de résilier ou non au plus tôt à partir de novembre.

Avant la loi de résiliation infra-annuelle

Les assurés ne peuvent que résilier à leur échéance annuelle, en adressant une lettre recommandée à l'organisme d'assurance au moins deux mois avant la date d'échéance.
En conséquence, l'observation de l'effet de résiliation de l'indexation tarifaire est entre 2 envois de courriers d'indexation.

La figure 2.2 schématise le principe de l'architecture de notre base.

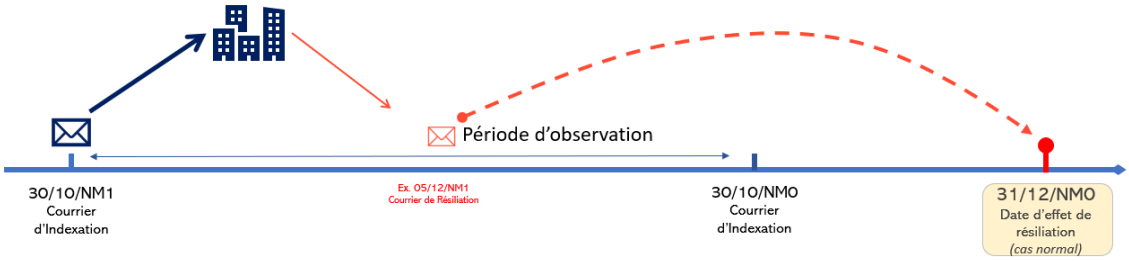

FIGURE 2.2 – Illustration de la fenêtre d'observation avant la loi de résiliation infra-annuelle

Prenons ainsi l'exemple d'un assuré A, il adresse son courrier de résiliation 05/12/NM1, le courrier n'est pas envoyé au moins deux mois avant la date d'échéance, donc la résiliation n'est pas appliquée pour l'année NM1, dans ce cas-là, la résiliation prendra effet le 31/12/NM0.

Après la loi de résiliation infra-annuelle

La loi dite Résiliation infra-annuelle donnera la possibilité aux assurés de contrats collectifs de résilier à tout moment leur contrat de complémentaire santé, une fois la première année de souscription passée.
Avant la mise en place de RIA, pour analyser l'effet de l'indexation sur la résiliation, il faut attendre de quelques mois à un an. Ainsi cette loi nous permet de mieux analyser l'impact de l'indexation sur la résiliation.
La figure 2.3 schématise le principe de l'architecture de notre base.

FIGURE 2.3 – Illustration de la fenêtre d'observation après la loi de résiliation infra-annuelle

Prenons ainsi l'exemple une entreprise assurée A, il adresse son courrier de résiliation 05/12/NM1, la résiliation prendra effet le 31/12/NM1. Un assuré B réalise que sa cotisation est trop élevée 5 mois après avoir

été informé de son indexation, il adresse son courrier de résiliation 14/04/NM0, la résiliation prendra effet le 31/12/NM0.

2.3.2 Choix des motifs de résiliation pertinents

Selon les cas, les contrats ne sont pas résiliés pour les mêmes raisons. En 2021, un pourcentage non-négligeable des contrats en portefeuille a été résilié, parmi lesquels la moitié ne répondait pas à une augmentation du tarif, mais plutôt à des circonstances comme démission initiative entreprise, ou encore la décision de l'assureur de résilier ce contrat pour cause de non-paiement de prime.

En matière de résiliations, il s'agit de déterminer si cette résiliation est liée à l'indexation ou non, et si elle est liée à l'indexation (par exemple : une résiliation pour cause de Fermeture définitive contrat santé n'est pas liée à l'indexation, alors une résiliation "Révision tarifaire" sera considérée comme raison liée à une hausse d'augmentation tarifaire). Nous considérerons que la résiliation est liée à l'indexation si elle apparaît dans le périmètre.

2.3.3 Autres critères pour préciser le périmètre

Afin d'avoir un résultat robuste, un dernier point est d'affiner notre périmètre d'étude. Nous ne prendrons en compte que les entreprises qui ont :

- au moins 12 mois d'ancienneté,
- 50 000 euros > montant de cotisation > 300 euros, parce qu'une fois que la cotisation dépasse le montant de 50 000, les processus de renouvellement annuel sont différents,
- le taux d'indexation >= 0,
- le minimum date début contrat > année de l'observation

Finalement, notre étude se porte sur les contrats collectifs de santé. Parmi ces contrats, ce qui nous intéresse, ce sont des entreprises qui sont passées par le processus de l'indexation, on exclut également les motifs de résiliation autre que l'indexation tarifaire telle que la cessation d'activité et non-paiement des cotisations, ainsi le périmètre d'étude se limite sur les entreprises non résiliées et les entreprises résiliées pour la causalité d'indexation tarifaire sur la profondeur de 2019-2021.

2.4 Préparation des données

Une étude statistique nécessite obligatoirement des pré-traitements sur les données afin de ne pas être perturbé par les points suivants :

- des incohérences dans les données,
- des valeurs aberrantes,
- des valeurs manquantes.

En fonction des problèmes rencontrés, plusieurs solutions se présentent. Si une variable possède trop de valeurs manquantes, il peut être préférable de la supprimer. Si des lignes présentent des incohérences, il est de même possible de les supprimer si celles-ci ne sont pas trop nombreuses. Dans les autres cas, si supprimer la ligne ou supprimer la variable n'est pas souhaitable, il est nécessaire d'effectuer des retraitements sur les données.

2.4.1 Contrôle de cohérence

Afin de vérifier la fiabilité de la base de données construite, on cherche à nettoyer la base de portefeuille en utilisant des tests de cohérence :

Vérification des valeurs manquantes

Ainsi, les traitements décrits précédemment permettent d'obtenir une base de données constituée de 10 408 lignes et des 21 variables suivantes, présentées dans la figure 2.4 ci-contre selon le volume de données manquantes qui leur est associé :

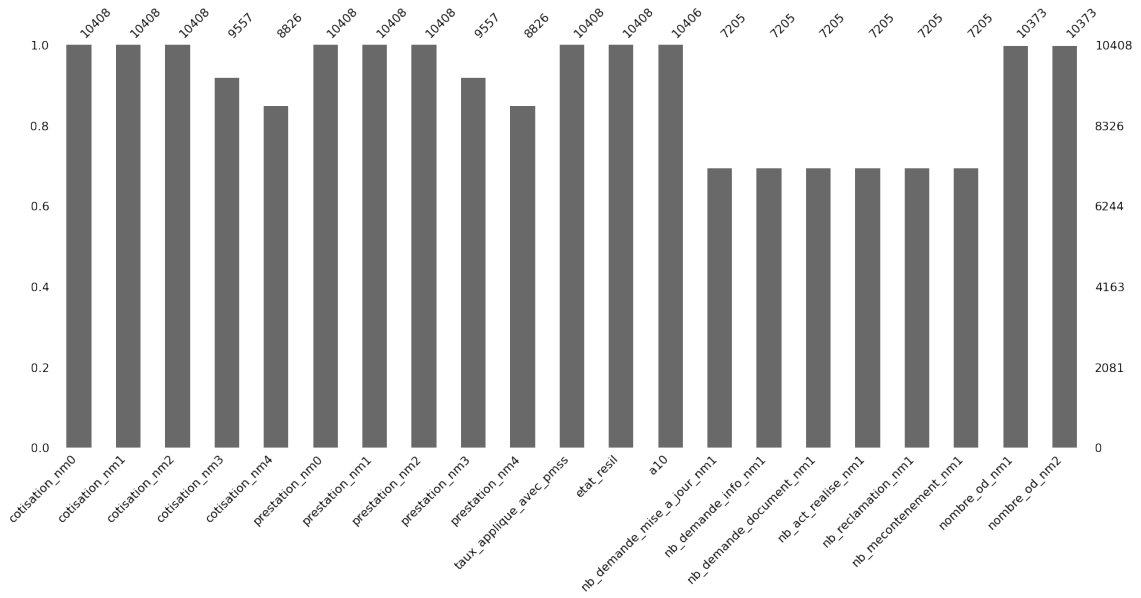

FIGURE 2.4 – Variables des bases fusionnées par proportion de données manquantes

Cette représentation permet de montrer que la base contient des lignes pour lesquelles les cotisations et les prestations ne sont pas renseignées. Il n'est pas envisageable de supprimer les lignes associées car cela induirait un biais trop important dans l'étude. Dans ce cas-là, les valeurs non renseignées sont représentées par zéro. Pour les autres variables, Les valeurs manquantes dans l'ensemble de données sont traitées en supprimant les lignes associées.

2.4.2 Création de nouvelles variables

Cette étape consistait à enrichir la base de données avec de nouvelles variables permettant une modélisation optimale ainsi qu'une étude statistique plus fine.

Les variables construites pour la modélisation

La variable ratio S/P brut est ajoutée à notre base de données :

$$\frac{S}{P} = \frac{\text{prestations (remboursement complémentaire)}}{\text{prime annuelle encaissée}}$$

Un indicateur du type de risque, prenant trois modalités différentes :

Bon risque	$0 < \frac{S}{P} \leq 80\%$
Risque modéré	$80\% < \frac{S}{P} \leq 100\%$
Mauvais risque	$100\% < \frac{S}{P}$

Les variables évolutions de cotisations et prestations sont ajoutées car il existe une forte corrélation entre les cotisations de différentes années.

$$\textbf{evol\_cotis\_NMi} = \frac{\textbf{Cotisation\_NMi}}{\textbf{Cotisation\_NMi+1}} - 1$$

$$\textbf{evol\_prest\_NMi} = \frac{\textbf{Prestation\_NMi}}{\textbf{Prestation\_NMi+1}} - 1$$

2.5 Présentation de la base de données (Analyse descriptive)

Il est important, avant de rentrer dans nos modèles, d'effectuer une étude descriptive qui nous permettait de mieux connaître nos variables, de détecter d'éventuelles tendances et informations importantes dans notre base de données, de vérifier si notre étude a du sens et un réel intérêt stratégique, ce qui permettra de mieux diriger notre étude par la suite.

2.5.1 Statistiques univariées

Répartition du nombre de résiliations d'entreprise par mois de survenance
Nous avons d'abord examiné la répartition du volume des courriers de résiliation sur douze mois de 2019 à 2021.

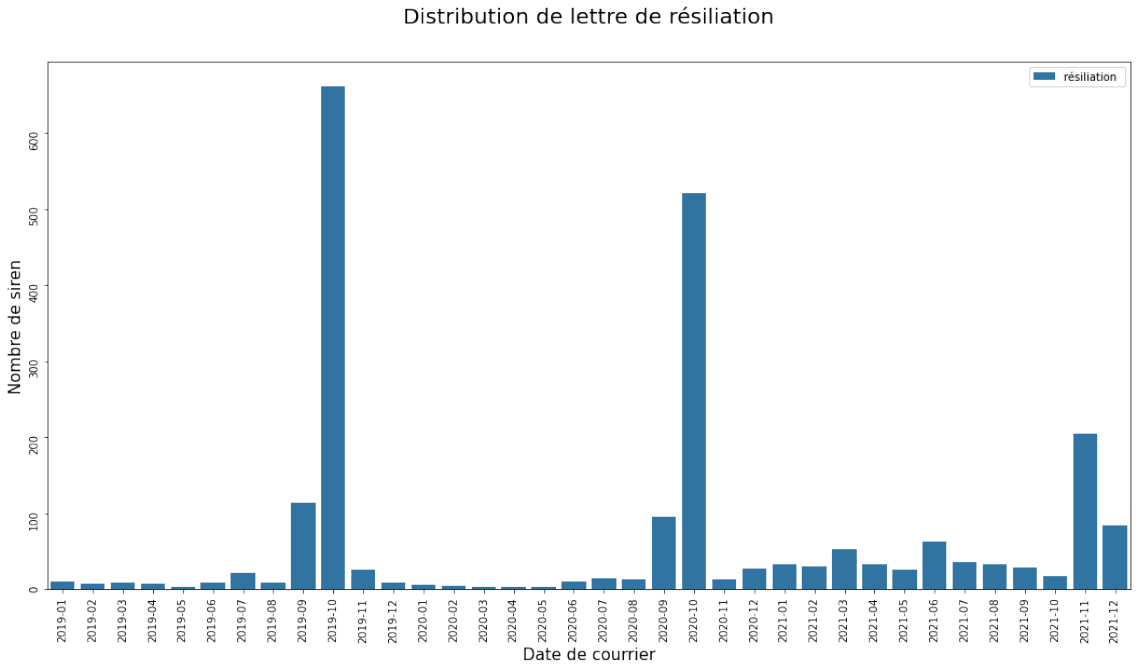

FIGURE 2.5 – Répartition du nombre de résiliation d'entreprise par mois

Avant la mise en place de la loi Résiliation Infra-Annuelle, les entreprises ont tendance à envoyer leur lettre de résiliation vers le mois d'octobre, ce qui explique l'augmentation des flux de courriers de résiliations vers le mois d'octobre. La mise en place de résiliations infra-annuelles dans le secteur de la complémentaire santé a changé le comportement de l'assuré. On constate une hausse de nombre de lettres de résiliation de mois novembre à août et une diminution de nombre de courriers de résiliation pendant le mois d'octobre et le mois de septembre.

La distribution de taux d'indexation

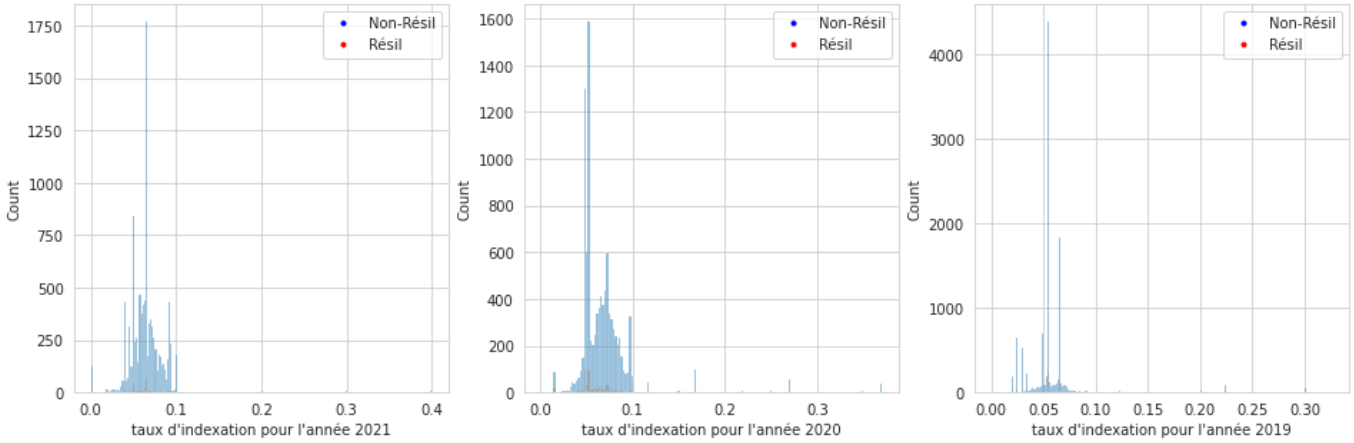

FIGURE 2.6 – Représentation de la répartition du taux d'indexation de 2019 à 2021

Dans cet histogramme des taux d'indexation, le pic des données survient autour de 0.05. La dispersion des données s'étend de 0 à 10 % environ. Pour l'année 2019 et 2020, sauf la valeur autour de 0.05, les données sont asymétriques. Nous remarquons également que certaines entreprises ont reçu un taux plus de 0.1 %.

Déséquilibre de variable cible

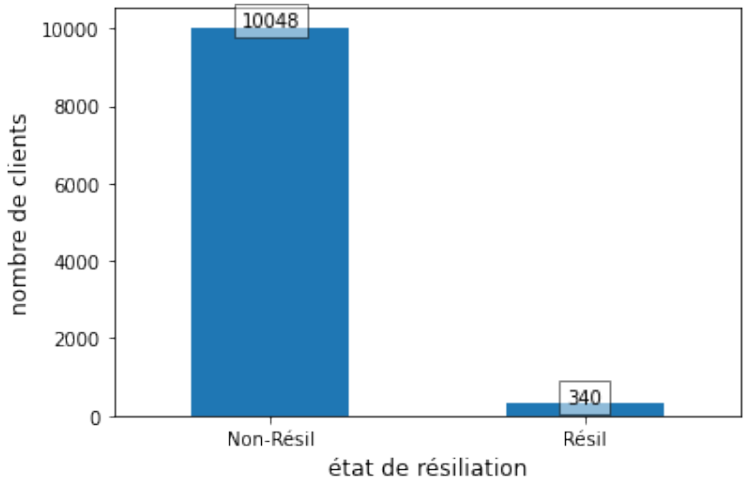

FIGURE 2.7 – Représentation de la répartition du nombre d'entreprises assurées

La base d'étude présente 10 388 entreprises durant l'année d'observation 2021, 340 résiliations sont observées. Le taux moyen de résiliation assuré est donc de 3,3 % dans notre périmètre d'étude. L'ensemble de données est déséquilibré, dans ce cas, cela signifie que la majorité des clients sont actifs, tandis que la minorité est constituée de clients qui résilient.

Cette faible fréquence des résiliations constitue un obstacle à l'apprentissage d'une classification binaire car les Non-Résiliés ont tendance à être privilégiés par la majorité des algorithmes. Le traitement du déséquilibre peut se faire de plusieurs manières différentes ; dans ce projet, plusieurs techniques de ré-échantillonnage ont été utilisées pour atteindre des proportions égales des deux classes.

2.5.2 Statistiques multivariées : évolution du taux de résiliation

Dans cette section, on se propose de faire des statistiques multivariées sur certaines variables clés par rapport à la résiliation. Puisqu'une variable explicative continue n'est pas nécessairement linéaire rapport à la variable à expliquer, elles ont été discrétisées. On cherche à diviser la variable en plusieurs tranches, puis les regrouper en fonction de ceux présentant la même mesure de risque. C'est-à-dire des tranches selon lesquelles les agrégations ne modifient pas statistiquement les mesures.

Notons que le taux de résiliation au global, c'est-à-dire le rapport entre nombre d'entreprises résiliées et le nombre total d'entreprises du portefeuille est de 3,3 %.

On définit le taux de résiliation comme suivante :

$$\text{taux de résiliation} = \frac{\text{nombre de siren résil}}{\text{nombre de siren total}}$$

Les graphiques ci-dessous représentent respectivement l'évolution du taux de résiliation suivant la variation de plusieurs variables : cotisation, taux d'indexation, S sur P, secteur d'activité. Les barres jaune foncé représentent le nombre d'entreprises totales, les barres jaune claire représentent le nombre d'entreprises résiliées. Les courbes bleues représentent le taux de résiliation que nous voulons observer.

Impact de l'indexation
Compte tenu du sujet de notre étude, la première dimension à étudier est le taux de l'indexation : comment évolue le taux de résiliation en fonction du taux de l'indexation appliqué à chaque entreprise. Avec la figure 2.8, nous constatons bien une augmentation du taux de l'indexation n'entraîne pas forcément une augmentation de taux de résiliation. Cependant, la forte indexation de plus de 10 % affecte le taux de résiliation qui grimpe rapidement. Ce qui nous parait aussi logique car les entreprises ont tendance à résilier leurs contrats si le taux est trop élevé.

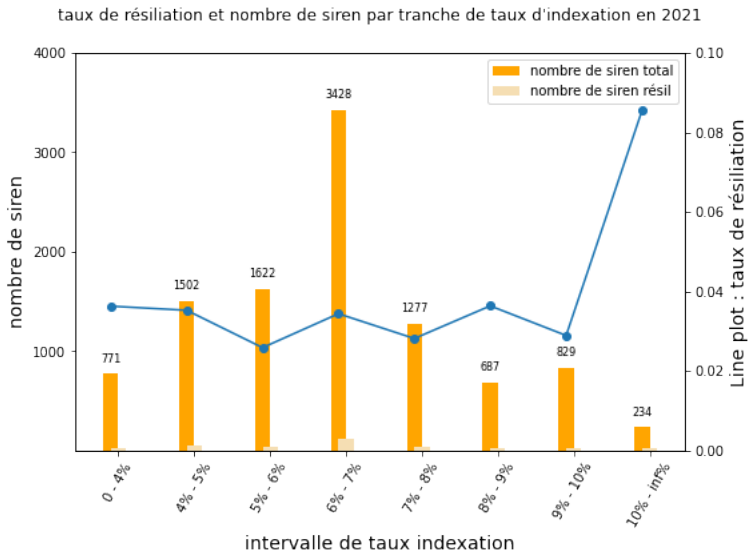

FIGURE 2.8 – L'évolution du taux de résiliation en fonction du taux d'indexation

Impact des primes
Notons que pour l'étude descriptive, la variable quantitative « Prime Brute Annuelle » a été découpée en classes pour plus de lisibilité. Les statistiques présentées pour estimer la liaison entre la prime et la variable

cible de résiliation ont aussi été calculées à partir de la variable découpée en tranches. Cette variable sera pourtant bien introduite sous sa forme continue dans les modèles de Machine learning.

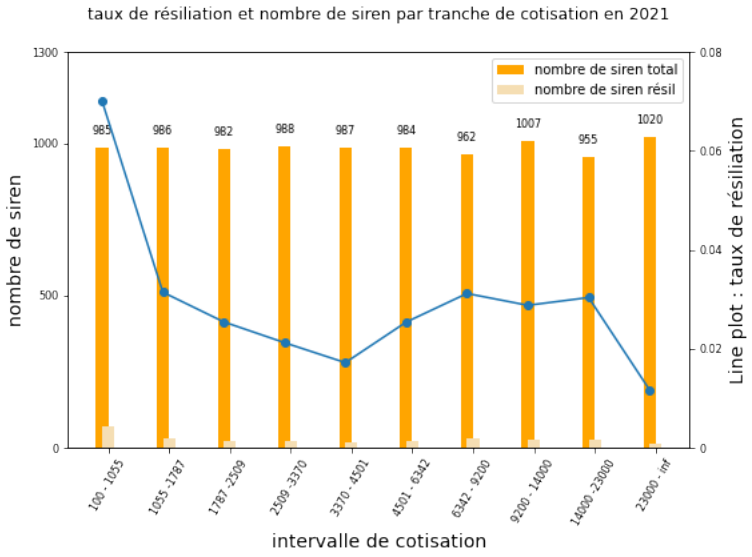

FIGURE 2.9 – taux de résiliation par niveau de prime en 2021

Le taux de résiliation de l'assuré semble avoir une évolution presque linéaire en fonction du niveau de la prime. On pourrait donc penser à modéliser le taux de résiliation en utilisant la variable cotisation.

Impact des S sur P brut

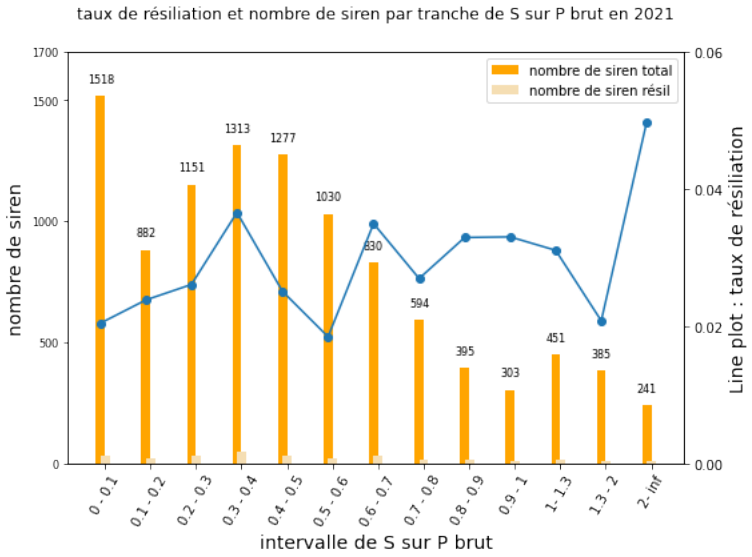

FIGURE 2.10 – taux de résiliation par tranche de S sur P en 2021

On remarque une légère tendance linéaire entre l'intervalle de S sur P et le taux de résiliation.(Si on exclut le S sur P supérieur à 1). Les taux de résiliation sont fortement différenciés par l'intervalle de S sur P.

Nous pouvons constater que les résiliations chez les « bons risques » sont moins nombreuses que chez les mauvais, puisque les taux de résiliation évoluent avec le S/P.

Le secteur d'activité
On utilisera le graphique 2.11 présentant le taux de résiliation en fonction du secteur.
On rappelle les différentes modalités :

Niveau d'agrégation de la NA10	Secteurs d'activité agrégés Intitulé court
BE	Industrie
FZ	Construction
GI	Commerce, transports, hébergement et restauration
JZ	Information et télécommunication
KZ	Activités financières
LZ	Activités immobilières
MN	Activités de services
OQ	Enseignement, santé, action sociale
RU	Autres activités de services

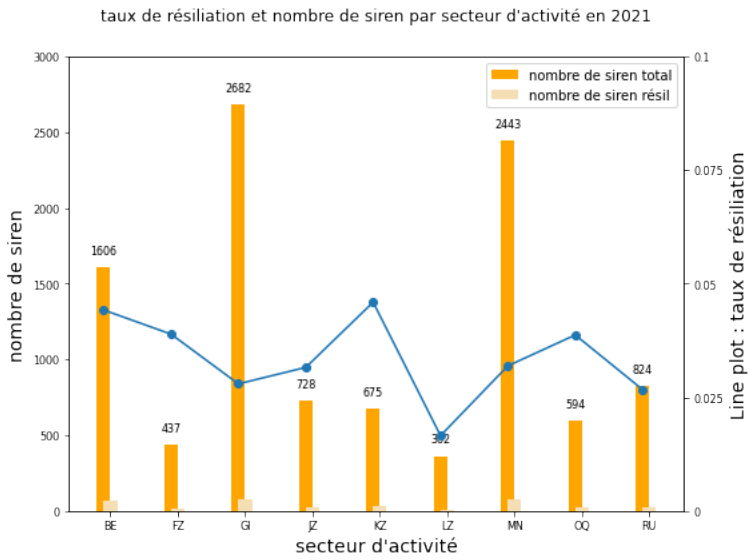

FIGURE 2.11 – taux de résiliation par secteur d'activité en 2021

Il y a une sur-représentation des secteurs Commerce, transports, hébergement et restauration, Activités de services, Industrie. Parmi les secteurs d'activité les plus impactés par la résiliation, l'industrie et le milieu financier passent pour des évidences.

2.5.3 Mesure de la corrélation

L'étude de la corrélation des variables est une étape importante avant de commencer la modélisation. Si deux variables sont entièrement corrélées ou anti-corrélées (coefficient de corrélation de Spearman égal à 1 ou -1), alors il y a une redondance d'information et cela risque de créer un déséquilibre lors de l'apprentissage. Le coefficient de Spearman est fondé sur l'étude de la différence des rangs entre les attributs des individus pour deux variables. Ce coefficient varie entre -1 et 1.

La figure 2.12 présente les corrélations entre les variables. Étant donné que les variables cotisations mesurent la même chose, mais sur des périodes différentes, il est intuitif que ces variables soient fortement corrélées (idem pour les prestations). Ceci est également visible dans le graphique, avec les corrélations élevées. On peut également conclure que les années proches les uns des autres sont encore plus corrélées. Le fait d'avoir une corrélation positive autour de 0.8 entre les variables nombres d'ouvrant droit et les cotisations, signifie que plus le nombre d'ouvrant droit augmente, plus la cotisation augmente.

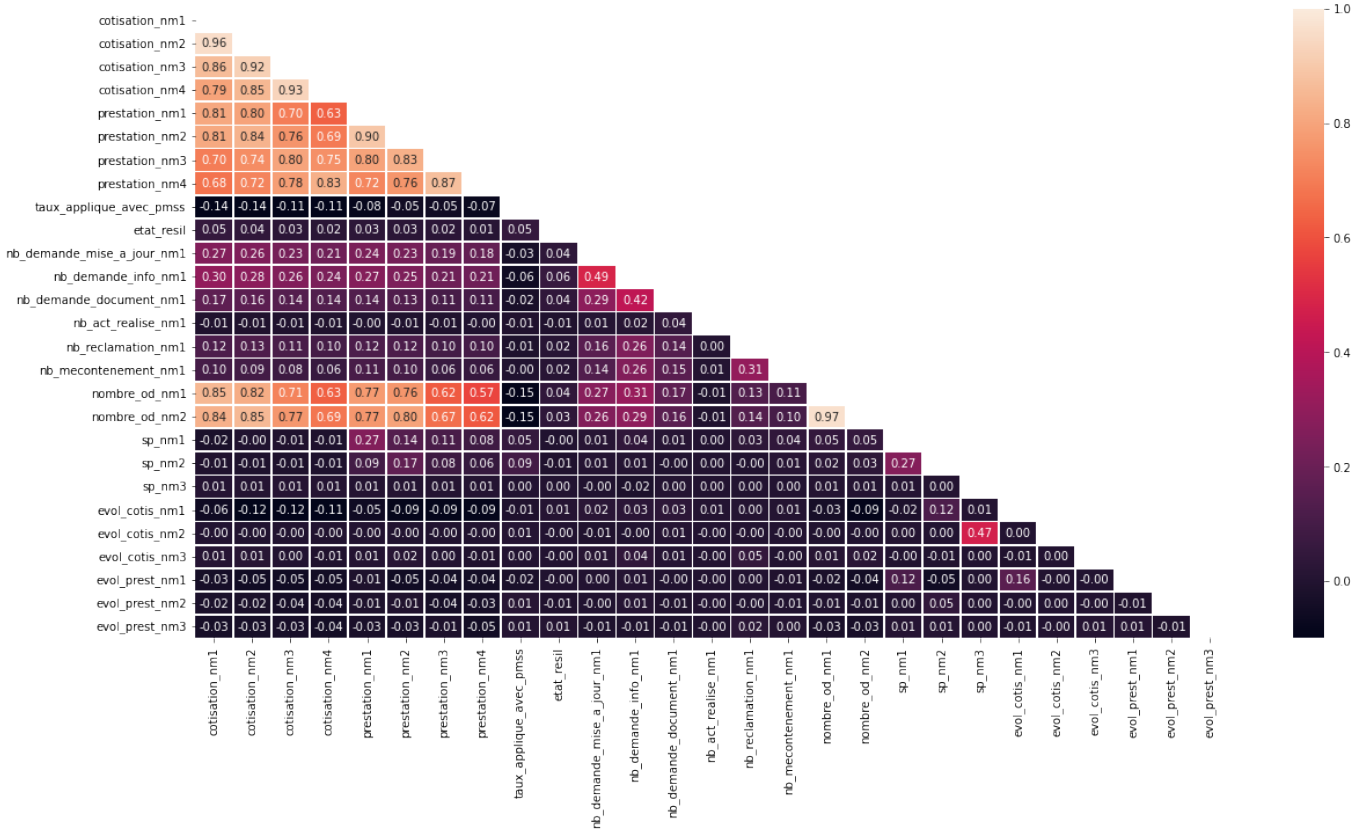

FIGURE 2.12 – Corrélation entre les différentes variables.

De ce fait, on remarque d'une part une dépendance modérée, faible, entre la variable taux d'indexation et les autres variables.

Conclusion du chapitre

L'analyse exploratoire de la base d'étude a montré que la plupart des variables explicatives présentes sont suffisamment liées à la variable cible de résiliation pour être introduites dans la modélisation.

Les statistiques descriptives montrent que le niveau d'indexation tarifaire a un impact sur le taux de résiliation : plus la prime est importante, plus la probabilité qu'un client résilie son contrat augmente. L'analyse descriptive révèle également que, suivant leur profil et les caractéristiques de leur contrat, les clients présentent des différents comportements. Cela indique que notre étude a bien un intérêt stratégique : pour retenir au mieux les clients, la marge de manœuvre de l'assureur réside, du moins en partie, dans l'établissement d'une indexation adéquate en fonction du client et de son contrat.

Chapitre 3

L'approche théorique

Nous souhaitons construire un modèle qui prédit la probabilité de résiliation pour chaque assuré. La variable à expliquer se décline donc en deux modalités :
– l'assuré résilié : Y = 1
– l'assuré non-résilié : Y = 0

L'outil classique pour répondre à ce genre de problème est l'utilisation des modèles issus de l'apprentissage automatique. Après quelques généralités sur l'apprentissage automatique supervisé, cinq méthodes de classification vont être présentées dans ce chapitre respectivement la régression logistique, l'arbre de régression, la forêt aléatoire, gradient boosting et machine à vecteur support. Dans la suite, nous introduirons les outils d'évaluation de la performance de modèle. En apprentissage, le déséquilibre des classes peut conduire à une mauvaise performance des classificateurs. Nous avons choisi de travailler sur quatre techniques permettant de corriger le déséquilibre des classes à savoir le ré-échantillonnage de la base. Enfin, nous présenterons la notion de l'élasticité.

3.1 Machine Learning

L'apprentissage automatique est un domaine qui se concentre sur la construction d'algorithmes qui font des prédictions à partir de données. Une tâche d'apprentissage automatique vise à identifier une fonction f : X → Y qui fait correspondre le domaine d'entrée X (des données) au domaine de sortie Y (des prédictions possibles). Les fonctions f sont choisies parmi différentes classes de fonctions, en fonction du type d'algorithme d'apprentissage utilisé.

Au cours des 20 dernières années, une variété de différentes techniques d'apprentissage automatique a été développées. Toutes ces techniques Toutes peuvent être divisées en deux catégories : supervisées et non supervisées, selon qu'elles disposent ou non d'instances étiquetées. Si c'est le cas, ces techniques sont dites supervisées, sinon elles sont dites non supervisées. Dans la présente étude, seul le premier type est pertinent et sera pris en considération.

3.1.1 Apprentissage automatique supervisé

L'apprentissage supervisé comprend toutes les tâches pour lesquelles l'algorithme a accès à des valeurs d'entrée et de sortie. Les valeurs d'entrée sont définies comme les informations externes que l'algorithme est autorisé à utiliser, tandis que les valeurs de sortie sont les étiquettes spécifiques de l'attribut de classe. Cela signifie que la structure des données est déjà connue et le but de ces programmes est d'assigner de nouvelles données aux bonnes classes.

3.1.2 Validation croisée k-fold

L'approche de validation croisée à k plis (aléatoire) consiste à répartir la base de données en k sous ensembles, puis d'en choisir un parmi ces k sous ensembles pour le test. Les $k-1$ restant sont utilisés pour l'apprentissage. Cette opération est répétée k fois afin que chaque sous ensemble soit utilisé une fois pour le test, voir l'illustration de la figure 3.1. L'erreur quadratique moyenne (EQM) est ensuite calculée pour le pli laissé de côté et répétée k fois, différentes observations sont utilisées pour la validation. L'EQM sera calculée pour chaque pli et l'estimation de la validation croisée k fois est la moyenne de celles-ci.

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

où $\hat{f}\left(x_i\right)$ est la prédiction que $\hat{f}$ donne pour la ième observation. L'estimation du k-fold $\mathrm{CV}_{(k)}$ est calculée en faisant la moyenne de ces valeurs par l'équation suivante :

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

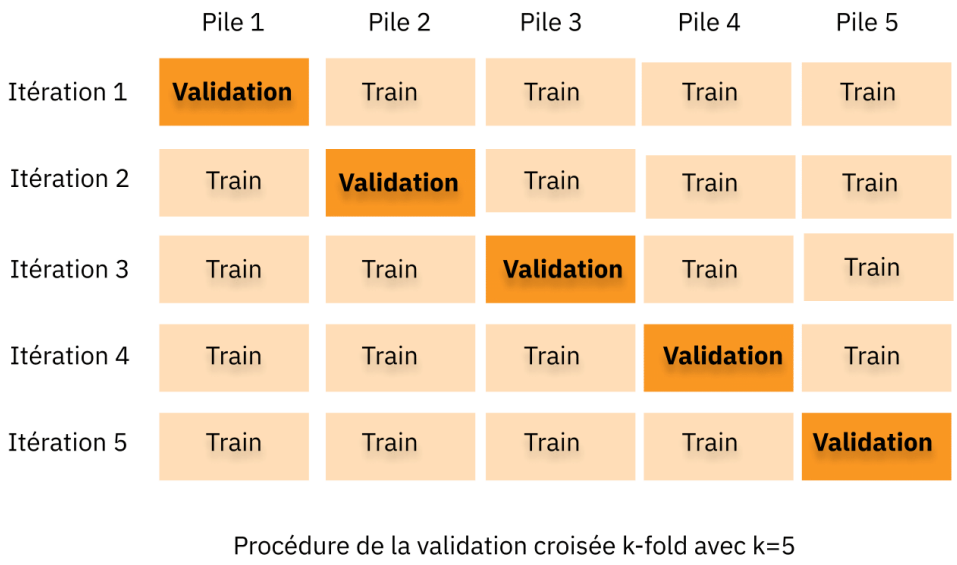

FIGURE 3.1 – Procédure de la validation croisée k-fold avec k=5

3.1.3 Grid Search

Le Grid Search croise simplement chacune de ces hypothèses et va créer un modèle pour chaque combinaison de paramètres. La méthode consiste à découper la data set en k échantillons. On sélectionne x échantillons pour constituer l'échantillon d'apprentissage. Les k-x échantillons restants permettront d'évaluer la performance du modèle. Pour construire le modèle suivant on sélectionne les échantillons différemment de manière à ne jamais avoir les mêmes échantillons d'apprentissage et de validation.

Une fois que chaque modèle a pu être entraîné et évalué, il ne reste plus qu'à comparer la performance pour choisir le meilleur modèle.

3.2 Algorithmes d'apprentissages

3.2.1 Régression logistique

On considère une population $\mathscr{P}$ divisée en 2 groupes d'individus G_1 et G_2 distinguables par des variables $X_1, \ldots, X_p$. Soit Y la variable qualitative valant 1 si l'individu considéré appartient à G_1 et 0 sinon. On souhaite expliquer Y à partir de $X_1, \ldots, X_p$.

On souhaite estimer la probabilité inconnue qu'un individu vérifiant $\big(X_1, \ldots, X_p\big) = x$ appartiennent au groupe G1 :

$$p(x) = \mathbb{P}\big(Y = 1 \mid (X_1, \ldots, X_p) = x\big), \quad x = (x_1, \ldots, x_p),$$

La probabilité $p(x)$ est aussi la valeur moyenne de Y quand $p(x) = \mathbb{E}\big(Y \mid ((X_1, \ldots, X_p) = x\big)$.

Si on exprime $p(x)$ avec $x = \big(x_1, \ldots, x_p\big)$ comme $p(x) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$, où p est le nombre de prédicteurs et β_i est le coefficient de l'observation correspondante X_i, alors au moins 2 problèmes surviennent :

- on a $p(x) \in [0,1]$ alors que $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \notin [0,1]$ a priori,
- quand $p(x)$ tend vers 0 ou 1, on doit avoir $\frac{\partial}{\partial x_j} p(x)$ qui tend vers 0. Or $\frac{\partial}{\partial x_j} p(x) = \beta_j$ ne tend pas vers 0 a priori.

Donc une fonction de lien est utilisée pour modéliser la probabilité que la réponse Y appartienne à une classe spécifique. Ici, nous utilisons la fonction logit :

$$\text{logit}(y) = \ln\left(\frac{y}{1-y}\right) \in \mathbb{R}, \quad y \in (0,1[.$$

Son inverse est la fonction :

$$\text{logit}^{-1}(y) = \frac{\exp(y)}{1+\exp(y)} \in (0,1[, \quad y \in \mathbb{R}.$$

Donc, on a

$$\text{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \quad x = \big(x_1, \ldots, x_p\big),$$

où $\beta_0, \ldots, \beta_p$ sont des coefficients réels inconnus.

Ainsi, $p(x)$ et $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ sont liés par la transformation logit. On en déduit l'expression de $p(x)$:

$$p(x) = \text{logit}^{-1}\big(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\big) = \frac{\exp\big(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\big)}{1 + \exp\big(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\big)}$$

On souhaite estimer $\beta_0, \ldots, \beta_p$ à partir des données, ce qui amènera une estimation de p(x) par substitution. Pour ce faire, on utilise la méthode du maximum de vraisemblance.

3.2.2 Les arbres de classification

Un arbre de décision est constitué de nœuds et de feuilles. Chaque nœud représente un attribut et chaque feuille correspond à une classe. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables

qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille.

Supposons un échantillon $(X_i, Y_i)_{i=1,\dots,n}$ avec $X_i \in \mathbb{R}^d$ les variables explicatives et $Y_i \in \mathbb{R}$ la variable de réponse. La construction de l'arbre consiste à déterminer une séquence de noeuds qui passe par une définition d'un critère de division optimal, une règle de décision du noeud final et un critère d'affectation de chaque feuille obtenu à une valeur de la variable de réponse.

Critère de division
La construction de l'arbre passe par une construction des différentes branches constitutives, lesquelles dépendent de la nature des variables explicatives. Si la variable explicative est qualitative ordinale ou quantitative à k valeurs, alors nous avons $(k-1)$ branches admissibles [1]. Dans le cas d'une variable à k modalités, nous avons $2^{(k-1)}-1$ branches admissibles. Le critère de division repose sur la définition d'une fonction d'hétérogénéité. Supposons que la variable à expliquer qualitative Y a m modalités $T_1, \dots, T_m$, nous définissons la probabilité qu'un élément du j-ième noeud appartienne à la i-ème classe par :

$$p_{ij} = \mathbb{P}\left(T_i \mid \text{ Classe } j\right) \text{ avec } \sum_{i=1}^{m} p_{ij} = 1$$

Les probabilités conditionnelles sont soit définies par la formule de Bayes lorsque les probabilités d'appartenance à une classe donnée est connue, soit estimées par des rapports d'effectif : $p_{ij} = \frac{n_{ij}}{\sum_{i=1}^{in} n_{ij}}$.

Pour choisir la meilleure caractéristique, i.e, la variable la plus discriminante, nous calculons le gain d'information de chacune en utilisant un critère de partitionnement. Une fonction couramment utilisée est l'indice de Gini. Il est défini comme suit :

$$G = 1 - \sum_{i=1}^{n} p_{ij}^2$$

et il mesure la variance totale de toutes les classes j. L'indice de Gini (ou impureté de Gini) mesure le degré ou la probabilité qu'une variable particulière soit mal classée lorsqu'elle est choisie au hasard. Les arbres de décision donnent souvent de bons résultats sur l'ensemble d'apprentissage, mais ils sont susceptibles d'avoir une variance élevée et d'être surajustés.

3.2.3 Forêts aléatoires

Une forêt aléatoire se compose de nombreux arbres de décision. Chaque arbre de décision dans la forêt est créé à partir d'échantillons bootstrap sélectionnés au hasard, qui sont agrégés pour faire une prédiction appelée "Bagging". Cette étape permet la diversité des forêts, ce qui entraîne une amélioration substantielle des performances de sa précision. Plus il y a d'arbres dans la forêt, plus la prédiction est robuste. Dans les forêts aléatoires, nous cultivons plusieurs arbres au lieu d'un seul arbre dans le modèle pour classer un nouvel objet. Sur la base des attributs, chaque arbre donne une classification et la forêt choisit la classe avec le plus de votes comme classificateur. Dans le cas de la régression, il prend la moyenne des sorties par différents arbres.

Bagging
La méthode du Bagging a été introduite par Breiman (1996). Le mot Bagging est la contraction des mots Bootstrap et Aggregating. Étant donné un échantillon d'apprentissage D, son idée de base est de générer, à partir d'une base d'apprentissage D, un ensemble de classifieurs différents et de les agréger ensuite dans un seul à l'aide un vote majoritaire par exemple. Dans le cas des forêts aléatoires, Breiman a utilisé la notion du bootstrap pour construire des sous ensembles d'exemples à partir de la base d'apprentissage D. Chaque sous ensemble D_i contient des exemples qui sont tirés aléatoirement et avec remise et donne naissance à un classiffieur qui est un arbre de décision.

Les défauts du bagging :
Une moyenne de B i.i.d. variables aléatoires, chacune avec une variance σ^2, ont une variance $\frac{1}{B}\sigma^2$. Si les variables sont simplement i.d. (identiquement distribué, mais pas nécessairement indépendant) avec une corrélation positive ρ, la variance de la moyenne est

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

À mesure que B augmente, le deuxième terme disparaît, mais le premier demeure. L'idée dans les forêts aléatoires est d'améliorer la réduction de la variance du bagging en réduisant la corrélation entre les arbres, sans trop augmenter la variance.

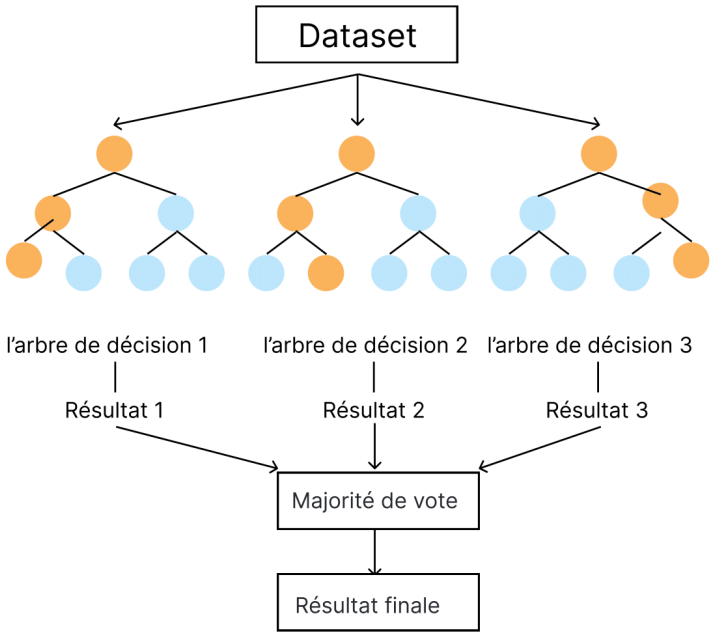

FIGURE 3.2 – Forêt aléatoire

Les forêts aléatoires sont donc une amélioration du bagging pour les arbres de décision CART dans le but de rendre les arbres utilisés plus indépendants (moins corrélés).

L'algorithme de forêt aléatoire

1. Pour $b = 1$ à B :

   (a) Prenons un échantillion bootstrap $\mathbf{D}^*$ de taille N à partir des données d'apprentissage.

   (b) Développez un arbre de forêt aléatoire T_b à partir de données bootstrap, en répétant de manière récursive les étapes suivantes pour chaque nœud terminal de l'arbre, jusqu'à ce que la taille de nœud minimale $n_{\min}$ soit atteinte.

      i. Sélectionnez m au hasard parmi les variables p.

ii. Choisissez la meilleure variable/split-point parmi les m.

iii. Divisez le nœud en deux nœuds filles.

2. Sortez l'ensemble d'arbres $\{T_b\}_1^B$.

Pour faire une prédiction à un nouveau point x :
Classification : Soit $\hat{C}_b(x)$ la prédiction de forêt aléatoire. Alors $\hat{C}_{\text{rf}}^B(x)$ = vote majoritaire $\left\{\hat{C}_b(x)\right\}_1^B$.

Feature Importance
La détermination de l'importance des caractéristiques peut se faire de différentes manières. Une variable est considérée comme importante si sa suppression entraîne une augmentation de l'erreur de prédiction. Une approche commune lorsqu'on considère Random Forest est de regarder l'indice de Gini. Chaque nœud dans l'algorithme Random Forest contient une distribution spécifique de la variable de réponse. Après chaque division, l'algorithme vise à rendre les nœuds enfants, c'est-à-dire les feuilles, aussi purs que possible. L'objectif est d'avoir des observations aussi similaires que possible dans chaque nœud. La pureté des nœuds peut être mesurée par l'indice de Gini, qui varie de 0 à 1.

3.2.4 Gradient Boosting

Le principe du boosting est de combiner les sorties de plusieurs classifieurs faibles (weak classifier) pour obtenir un résultat plus fort (strong classifier). Le classifieur faible doit avoir un comportement de base un peu meilleur que l'aléatoire : taux d'erreurs inférieur à 0.5 pour une classification binaire (c'est-à-dire qu'il ne se trompe pas plus d'une fois sur deux en moyenne, si la répartition des classes est équilibrée). Chaque classifieur faible est pondéré par la qualité de sa classification : mieux il classe, plus il sera important. Les exemples mal classés aurons un poids plus important (on dit qu'ils sont boostés) vis-à-vis de l'apprenant faible au prochain tour, afin qu'il pallie le manque. Cependant, il existe certaines restrictions du boosting.

Un des algorithmes les plus utilisés en boosting s'appele Gradient Boosting.

Soient Y la variable réponse et X = $\left(X_1, \ldots, X_p\right)$ les variables explicatives. Et soient $\left(y_i, x_i\right)_{1 \leq i \leq n}$ les n observations dont nous disposons. L'algorithme GBM peut être décrit comme suit :

1. Initialisation du prédicteur :
$$f_0(x) = \arg\min_c \sum_{i=1}^{n} L\left(y_i, c\right).$$
2. Pour m allant de 1 à M :
   (a) Pour i de 1 à n on calcule
   $$r_{im} = -\frac{\partial L\left(y_i, f_{m-1}\left(x_i\right)\right)}{\partial f_{m-1}\left(x_i\right)}$$
   (b) Appliquer un arbre de décision à la variable $r_m = (r_{im})_{1 \leq i \leq n}$. Cet arbre a J_m nœuds terminaux. On notera R_{jm} la région du jm-ième nœud terminal
   (c) Pour $jm = 1$ à J_m on calcule :
   $$\gamma_{jm} = \min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}\left(x_i\right) + \gamma\right).$$

(d) Mise à jour :

$$f_m(x) = f_{m-1}(x) + \delta \sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right), \quad \text{avec } \delta \epsilon [0,1]$$

Le δ est ce que l'on appelle le Learning Rate ou Shrinkage

(e) Le prédicteur final est donc

$$F(x) = f_M(x).$$

Nous constatons bien que le prédicteur d'initialisation du GBM est une constante. Elle est égale à la moyenne des observations. Nous remarquons aussi qu'à chaque étape m la base de modélisation devient $\left(\tilde{y}_{im}, x_i\right)_{1 \leq i \leq n}$
Avec :

$\tilde{y}_{im} = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}$, pour tout m > 0 Et $\tilde{y}_{i0} = y_i$.

Pour les problèmes de classification binaire la fonction de perte L usuelle est la déviance binomiale :

$$L(y,p) = y \log(p) + (1-y)\log(1-p).$$

3.2.5 Machine à vecteur support

Machine à vecteur support crée un ou plusieurs hyperplans dans un espace dimensionnel H. La première tentative dans le processus de division des données est toujours, d'essayer de séparer linéairement les données dans les étiquettes correspondantes. Donc on définit le séparateur f tel que :

$$f(x) = < \omega, x > + b, \quad \forall \omega \in H \text{ et } b \in \mathbb{R}$$

L'équation $f(x) = 0$ défini la frontière de séparation des deux classes. Il existe plusieurs hyperplans possibles pour séparer les classes (cf. figure de gauche ci-dessous) et il sera donc question de trouver celui qui parmi eux optimise au mieux la séparation des données. Lorsque $f(x) > 0$, le vecteur x appartient alors à la classe des échantillons dont l'étiquette est L et réciproquement lorsque $f(x) < 0$, le vecteur x appartient à la classe des échantillons d'étiquette K.

La variable à prédire est alors définie par la formulation :

$$Y = \begin{cases} +1 & \text{si } f(x) > 0 \\ -1 & \text{si } f(x) < 0 \end{cases}$$

En définissant la distance d'un point x_0 à l'hyperplan $H_{\omega,b} :< \omega, x > + b$ par : $d\left(x_0, H_{\omega,b}\right) = \frac{|\langle \omega, x > + b|}{\|\omega\|}$, le but des SVM sera de trouver l'hyperplan tel que sa distance aux points les plus proches soit maximale. Pour ce faire, les supports vectors étant les points x tel que $|\langle \omega, x \rangle + b = 1|$, la marge est définie par la distance des vectors supports à l'hyperplan et est égale à $\frac{2}{\|\omega\|^2}$. Afin d'obtenir l'hyperplan optimal, l'algorithme va minimiser l'inverse des marges sous contrainte que l'hyperplan $H_{\omega,b}$ sépare réellement les points :

$$\min \frac{1}{2} \langle \omega, \omega \rangle \text{ s.c } : \forall i, y_i \left(< \omega, x_i > + b\right) \geq 1$$

Ce programme d'optimisation s'avérant difficile à résoudre vu le nombre très important de contraintes, nous le résolvons dans l'espace dual pour diminuer la complexité du problème. Dans, l'espace initial nous avions $\omega = \sum \alpha_i y_i x_i$ avec $\sum \alpha_i y_i = 0$. Dans l'espace dual, ce programme s'écrit :

$$\min \left\{ \frac{1}{2} \alpha A^\top \alpha - 1^\top \alpha \right\} \text{ s.c. } \begin{cases} 0 \leq \alpha_i \forall i \\ y^\top \alpha = 0 \end{cases}$$

$$\text{Où } Q = \left[Q_{i,j}\right] \text{ et } Q_{i,j} = y_i y_j x_i^\top x_j$$

3.3 Les outils d'évaluation de la performance

Il existe plusieurs outils, applicables universellement, et permettant d'évaluer la performance de nos modèles et de les comparer. Nous nous limiterons aux critères les plus utilisés, qui sont définis par la suite.

Pour évaluer les performances du modèle sur l'ensemble de données déséquilibré, nous utilisons des métriques couramment utilisées telles que la matrice de confusion, la précision, le rappel, le score f1 et l'AUC.

3.3.1 La matrice de confusion

Une matrice de confusion est un tableau qui donne des informations sur la visualisation d'un modèle prédictif d'un algorithme ainsi que sur les classes qui sont anticipées avec précision, celles qui sont erronées et le type d'erreurs commises. Cette matrice compare les valeurs cible réelles avec celles prédites par le modèle.

Idéalement, un classifieur parfait nous donnerait 0 dans les deux cases d'erreurs (faux positifs et faux négatifs). Il est rare d'avoir un tel classifieur. En général, si on tente de diminuer la quantité de faux négatifs, le nombre de faux positifs augmentera et vice-versa. Pour cette raison, la matrice de confusion d'un classificateur sera obtenue pour un certain seuil d'erreur préalablement choisi. Simplement, en regardant la matrice de confusion, il est possible d'avoir une petite idée de la performance d'un classifieur. Il est toutefois possible de tirer plus d'informations de cette matrice.

	Classe prédite	
Classe réel	Résilié $Y = 1$	Non - Résilié $Y = 0$
Résilié	VP	FN
Non-Résilié	FP	VN

Vrai positif
La prédiction est positive et c'est la réalité.

Faux positif
La prédiction est positive et ce n'est pas la réalité.

Vrai négatif
La prédiction est négative et c'est la réalité.

Faux négatif
La prédiction est négative et ce n'est pas la réalité.

Accuracy
L'accuracy décrit la quantité relative de données correctement classées dans un ensemble de données. Si $\hat{y}_i$ est la prédiction pour le point de données i et y_i est l'étiquette réelle, le mauvais classement est défini comme suit :

$$accuracyc_n = \frac{1}{n} * \sum_i \left(y_i = \hat{y}_i\right)$$

Précision (taux de positifs prédits)
Cet indicateur représente la proportion d'éléments positifs bien prédits parmi tous les positifs observés et

se calcule par la formule $\frac{VP}{(VP+FP)}$

Recall (taux de vrais positifs)
Cet indicateur représente la proportion d'éléments positifs bien prédits parmi tous les positifs observés et se calcule par la formule $\frac{VP}{(VP+FN)}$

F1 mesure
Le F1-score permet de résumer les valeurs de la precision et du recall en une seule métrique. Mathématiquement, le F1-score est défini comme étant la moyenne harmonique de la precision et du recall, ce qui se traduit par l'équation suivante :

$$F1 = \frac{2 \times \text{ rappel } \times \text{ precision}}{\text{rappel+precision}}$$

3.3.2 La courbe ROC et le critère AUC

Une courbe ROC (Receiver Operating Feature Curve) est un graphique montrant les performances d'un modèle de classification à tous les seuils de classification. Cette courbe trace deux paramètres :

- Sensibilité : taux de vrais positifs
- Spécificité : taux de faux positifs

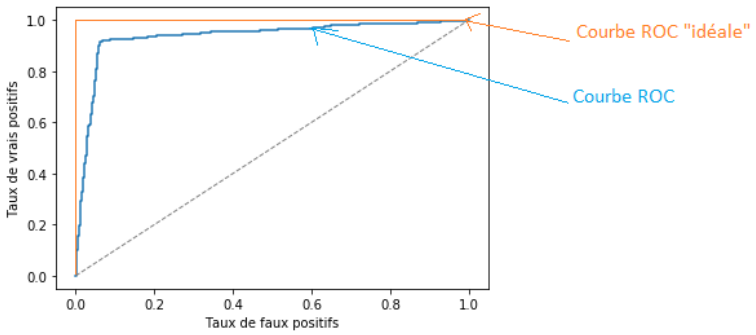

FIGURE 3.3 – Exemple de courbe ROC avec son AUC

La performance du modèle obtenu par la régression logistique peut être mesurée par l'analyse ROC (Receiver Operating Charactistic) . L'analyse ROC nous donne une courbe représentant la capacité du modèle à prédire la variable dépendante. Pour évaluer la performance des différentes méthodes présentées, nous utilisons en particulier la variable AUC (Area Under the Curve). La variable représente une proportion des valeurs de la variable dépendante ayant été prédites correctement par le modèle.

On peut représenter plusieurs courbes ROC de différents modèles sur le même graphe pour les comparer. Si les courbes se croisent alors on calcule l'AUC (Area Under the Curve), l'aire sous la courbe afin de pouvoir les comparer. Le modèle avec la plus grande aire sous la courbe correspond au plus performant.

Valeur de l'AUC	Interprétation
AUC = 0.5	Modèle aléatoire non discriminant
0.5 < AUC < 0.65	Discrimination médiocre
0.65 < AUC < 0.8	Bonne performance prédictive
0.8 < AUC < 0.9	Très bonne performance prédictive
AUC > 0.9	Modèle excellent

TABLE 3.1 – Synthèse des variables candidates

3.3.3 P value

Plusieurs variables sont collectées durant l'analyse de régression logistique. La p-value joue le même rôle que dans l'analyse de corrélation en représentant la probabilité d'obtenir un coefficient différent de zéro par chance

Hypothèse nulle, hypothèse alternative

L'hypothèse que l'on cherche à vérifier s'appelle hypothèse nulle (notée H_0) . Elle porte sur la loi de probabilité (ou de façon équivalente le paramètre θ de cette loi) ayant donné naissance à l'échantillon disponible.

$$H_0 : \theta \in \theta_0$$

Si le résultat d'échantillonnage conduit au rejet de l'hypothèse nulle, nous devrons alors en arriver à une autre conclusion. La conclusion acceptée dans ce cas s'appelle hypothèse alternative (notée H_1).

$$H_1 : \theta \in \theta_1$$

La p-value est la probabilité, sous H_0, d'obtenir une statistique aussi extrême (pour ne pas dire aussi grande) que la valeur observée sur l'échantillon. Aussi, pour un seuil de significativité α donné, on compare p-value et α, afin d'accepter, ou de rejeter H_0

- si $p \leq \alpha$, on va rejeter l'hypothèse H_0 (en faveur de H_1);
- si p > α, on va rejeter H_1 (en faveur de H_0).

On peut alors interpréter la p-value comme le plus petit seuil de significativité pour lequel l'hypothèse nulle est acceptée.

3.4 Données déséquilibrées et méthode d'échantillonnage

3.4.1 Définition

Imbalanced, mal balancé, skewed, ce type de problème de données est très fréquent. Il signifie qu'une classe dans un problème de classification est sous-représentée par rapport aux autres et que le modèle de machine learning n'est pas suffisamment pénalisé s'il n'en tient pas compte. Plusieurs auteurs s'accordent à dire que la situation de déséquilibre des classes est avérée à partir du moment où l'une des classes de la variable d'intérêt représente en proportion moins de 10 % des observations, pour d'autres le seuil retenu est plutôt de 5 %.
Le cas classique est un problème à deux classes, une majoritaire à 99 %, une minoritaire à 1 %. Un modèle qui répond toujours, la majorité est correcte 99 % du temps, mais il n'a rien appris puisque sa réponse est constante. Comment le forcer à apprendre quelque chose? La stratégie la plus couramment utilisée pour faire face au déséquilibre des classes est le ré-échantillonnage.

Modifier l'échantillon d'étude de façon à ce que le volume de réponses positives pour la variable d'intérêt soit plus important, permet d'obtenir un équilibre des classes de la variable d'intérêt.

3.4.2 Sur-échantillonnage

Le sur-échantillonnage est un moyen pour rééquilibrer les bases de données, il consiste à répliquer aléatoirement des individus appartenant à la classe minoritaire. On multiplie les exemples de la classe minoritaire de manière à lui donner plus de poids, ce qui peut entraîner un sur-apprentissage.

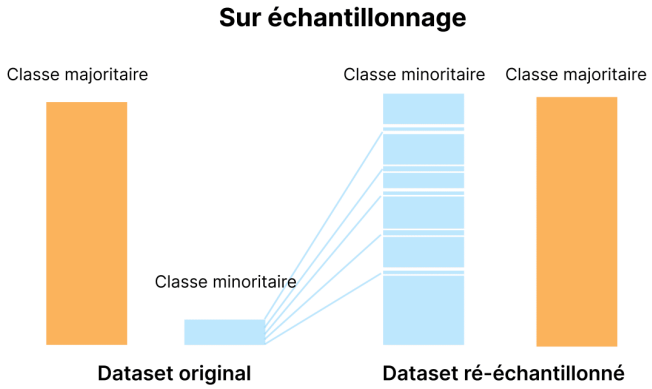

FIGURE 3.4 – Illustration de sur-échantillonnage

Il existe différentes méthodes de sur-échantillonnage. Nous couvrirons deux des plus populaires ci-dessous dans ce mémoire :

sur-échantillonnage aléatoire simple
L'approche de base de l'échantillonnage aléatoire avec remise à partir de la classe minoritaire.

Synthétique Minority Oversampling THechnique SMOTE
SMOTE est l'une des techniques de sur-échantillonnage les plus populaires développées par Chawla et al. (2002). Contrairement au sur-échantillonnage aléatoire qui ne duplique que certains exemples aléatoires de la classe minoritaire. SMOTE introduit les exemples synthétiques au hasard sur la ligne reliant le point de classe minoritaire concerné et un de ses K plus proches voisins. Selon le nombre de sur-échantillonnage requis les voisins des k plus proches voisins sont choisis aléatoirement, de sorte que les exemples générés sont différents de la classe minoritaire d'origine.

- Étape 1 : définition de l'ensemble de classes minoritaires A, pour chaque $x \in A$, les k plus proches voisins de x sont obtenus en calculant la distance euclidienne entre x et tous les autres échantillons de l'ensemble A.
- Étape 2 : le taux d'échantillonnage N est réglé en fonction de la proportion déséquilibrée. Pour chaque $x \in A$, N exemples (c'est-à-dire x1, x2, … xN) sont choisis au hasard parmi ses k plus proches voisins, et ils construisent l'ensemble A_1.
- Étape 3 : pour chaque exemple $x_k \in A_1$(k=1, 2, 3… N), la formule suivante est utilisée pour générer un nouvel exemple : $x' = x + rand(0,1) * \mid x - x_k \mid$ dans laquelle rand(0, 1) représente le nombre aléatoire entre 0 et 1.

3.4.3 Sous-échantillonnage

On réduit le nombre d'exemples de la classe majoritaire sans altérer la capacité du modèle à trouver une bonne solution, cela consiste à enlever des exemples loin de la frontière de classification. L'inconvénient de cette méthode est la perte d'information sur la classe majoritaire par rapport à la base initiale. On obtient une base certes équilibrée mais qui ne reflète pas toujours l'information contenue dans la base initiale.

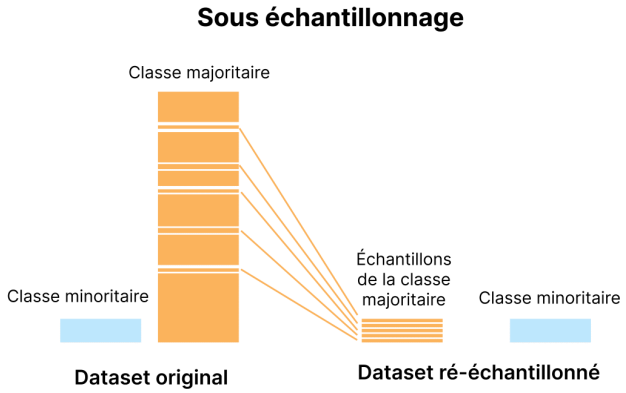

FIGURE 3.5 – Illustration de sous échantillonnage

Sous-échantillonnage aléatoire simple
Une technique simple de sous-échantillonnage consiste à sous-échantillonner la classe majoritaire de manière aléatoire et uniforme. Cela peut potentiellement entraîner une perte d'informations. Mais si les exemples de la classe majoritaire sont proches des autres, cette méthode pourrait donner de bons résultats.

Liens Tomek
Soit $d(x_i, x_j)$ la distance euclidienne entre x_i, x_j, où x_i désigne l'échantillon qui appartient à la classe minoritaire et x_j désigne l'échantillon qui appartient à la classe majoritaire. S'il n'y a pas d'échantillon, x_k satisfait la condition suivante :
1. $d(x_i, x_k) < d(x_i, x_j)$
2. $d(x_j, x_k) < d(x_i, x_j)$
alors la paire de (x_i, x_j) est un Tomek Link .

3.5 Notion d'élasticité

Nous utiliserons dès lors un terme élasticité. Il convient donc de l'expliciter afin d'en saisir réellement la signification. Dans un contexte économique, quand il s'agit d'un bien qui peut être consommé en quantités variables, l'élasticité du prix par rapport à la demande est définie comme suit :

$$E = -\frac{(Q_{P1} - Q_{P0}) / Q_{P0}}{(P1 - P0) / P0}$$

où

$E =$ l'élasticité de prix recherchée
$P0 =$ le prix du bien à la période 0
$P1 =$ le prix du bien à la période 1
$Q_{P0} =$ la quantité consommée au prix P0 par le consommateur
$Q_{P1} =$ la quantité consommée au prix P1 par le consommateur

L'élasticité permet d'évaluer la sensibilité de la quantité demandée à une variation du prix. Autrement dire :

$$E = \frac{\Delta Q/Q}{\Delta P/P} \rightarrow \frac{\partial Q}{\partial P} * \frac{P}{Q} \quad \text{when } \Delta \rightarrow 0$$

On utilise toujours la valeur absolue de l'élasticité-prix. Nous savons que la relation prix-quantité est négative pour la demande, mais c'est la valeur absolue qui nous intéresse.
On dit que la demande est :

$\|E\| > 1$	élastique
$\|E\| = 1$	élastique unitaire
$\|E\| < 1$	inélastique

L'élasticité nous permet donc de connaître de combien va varier la demande si nous modifions le prix. En effet, une élasticité égale à ε impliquerait que pour une hausse de 1% du prix, nous aurons une baisse de ε% de la demande.

Dans le cas présent, il n'y a pas de <quantité consommée> ; le comportement pertinent est la résiliation ou non du contrat par le client. Les comportements binaires (de la forme Oui / Non) sont analysés via leur probabilité d'occurrence. Il convient d'ajuster en conséquence la définition de l'élasticité prix.

Dans ce mémoire, l'élasticité prix soit définie : « La variation, en points de probabilité, de la probabilité estimée de résiliation du contrat quand le taux d'indexation appliqué varie d'un point de pourcentage ».

Chapitre 4

Modélisation de résiliation en cas de déséquilibre des classes

Notre but est de prédire le comportement du client au moment de renouvellement annuel, autrement dit au moment où le client réalise que la prime de son contrat a été majorée. Il faut donc modéliser le comportement du client, et, plus précisément, déterminer la probabilité qu'un client décide de mettre fin à son contrat lorsque ce dernier arrive à échéance. La résiliation dépendra bien entendu du profil de l'entreprise mais aussi des caractéristiques du contrat.

Dans ce chapitre, nous présentons les résultats obtenus sur les données origines déséquilibres et sur les données ré-échantillonnage pour les différents modèles construits. Nous disposons de plusieurs méthodes d'apprentissage tel que la régression logistique ou la forêt aléatoire, etc. Les modèles sont comparés afin de sélectionner le modèle qui a le meilleur F1-Score et maximise l'AUC. La procédure de modélisation utilisée peut être résumée comme la suivante :

1. Séparer 70 % des observations en base d'apprentissage et 30 % en test.

2. Sur la base d'apprentissage, on applique les cinq méthodes d'apprentissage, ensuite, la base de test sera utilisée pour tester les performances prédictives des classificateurs.

3. Quatre techniques de ré-échantillonnage vont être réalisées uniquement sur les données d'apprentissage, afin d'évaluer les performances des algorithmes, le recall est utilisé.

4.1 Résultat basé sur la base originale

Dans la section suivante, tout d'abord, le résultat de chaque modèle avec son meilleur ensemble d'hyperparamètres sera présenté ainsi qu'une matrice de confusion. L'importance des caractéristiques pour la forêt aléatoire et pour la régression logistique sera illustrée. Les modèles seront ensuite comparés et les performances seront évaluées à l'aide des métriques AUC, F1-score, précision et sensibilité.

Pour valider les différentes méthodes utilisées et calculer leurs performances, nous séparons la base de données en deux parties, un échantillon d'apprentissage et une base de test. La première base est constituée de 70 % des observations et servira à l'entraînement du modèle. Ces données permettront aux classificateurs d'apprendre des règles. La deuxième base contient le reste des données, soit 30 %. Cette base sera utilisée pour tester les performances prédictives des classificateurs et évaluer objectivement les erreurs réelles. Par conséquent, il n'est pas utilisé pour l'apprentissage, ce qui signifie que le modèle sélectionné est indépendant de cette base de test.

Base initiale apprentissage		Base initiale test	
Non-Résilié	Résilié	Non-Résilié	Résilié
7 017	242	3 031	98

TABLE 4.1 – Effectif des base d'apprentissage originale et base test originale

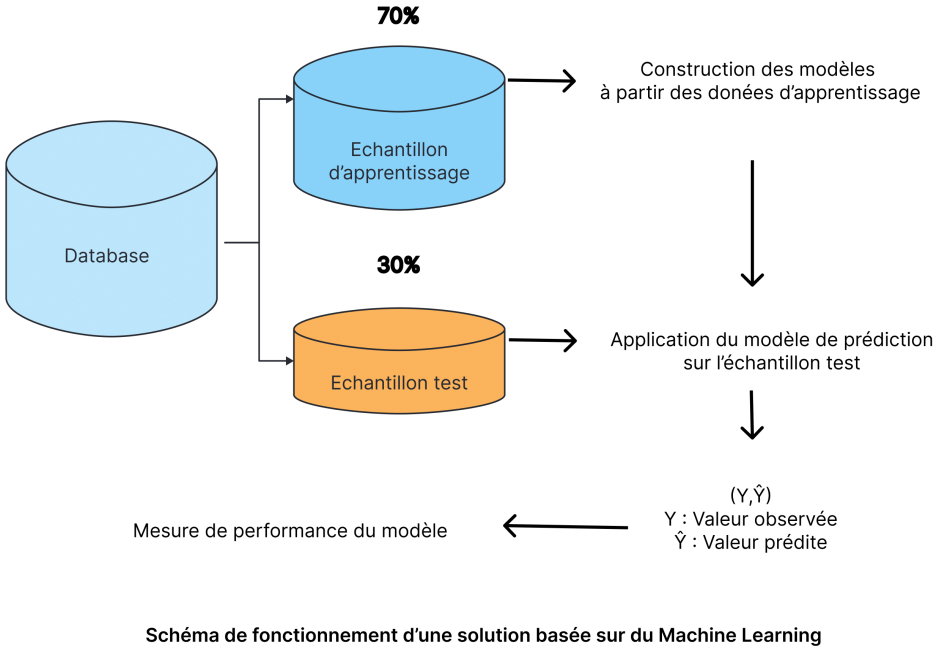

FIGURE 4.1 – Schéma de fonctionnement d'une solution basée sur du Machine learning

Après séparation de la base initiale, la composition de base est la suivante : La base de test est composée de 3 031 non-résilié, et 98 résilié, en total 3 129 entreprises.

4.1.1 Réglage des hyperparamètres

Afin d'ajuster le modèle et d'augmenter la puissance des prédictions, certains hyperparamètres ont été ajustés par grid-search, grid-search est une procédure permettant de trouver les paramètres optimaux pour un modèle donné. Pour chaque algorithme, ensuite, nous choisissons le meilleur ensemble de paramètres, à la fin de l'étape d'hyperparamétrage, nous obtenons un modèle avec des hyperparamètres optimaux.

4.1.2 Régression logistique

Nous venons de voir que l'étape de réglage des hyperparamètres a été effectuée pour les modèles.
Nous appliquons la base test sur le modèle régression logistique avec des hyperparamètres optimaux pour récupérer les mesures de performance et la matrice de confusion. Les premiers résultats de la régression logistique sont affichés ci-dessous :

la plupart des assurés sont dans la colonne des vrais négatifs, ce qui signifie que la majorité sont correctement classés comme actifs.

		Classe prédite		Total
		Résilié	Non-Résilié	
Classe réelle	Résilié	16	82	98
	Non-Résilié	135	2 896	3 031
Total		151	2 978	3 129

TABLE 4.2 – les résultats sur une matrice de confusion par régression logistique

Parmi les 151 individus prédits comme résiliés, il y a 16 bonnes prédictions, ce qui donne une **précision** de $\frac{16}{151} \approx 11\%$.

Sur les 98 assurés résiliés observés, 16 ont été bien prédits et 82 ont été mal classé. Cela donne une **sensibilité (recall)** de $\frac{16}{98} \approx 16\%$.

Le score F1 est généralement plus utile que la précision, surtout si nous avons une distribution de classe déséquilibrée. Il est la moyenne pondérée de la précision et du recall : $2 \times \frac{0.16 \times 0.11}{0.16+0.11} \approx 13\%$.

La matrice de confusion indique que nous avons $2896 + 16$ prédictions correctes et $135 + 82$ prédictions incorrectes, ce qui donne un taux d'erreur de $\frac{217}{3129} \approx 7\%$, et un accurary de 93 %.

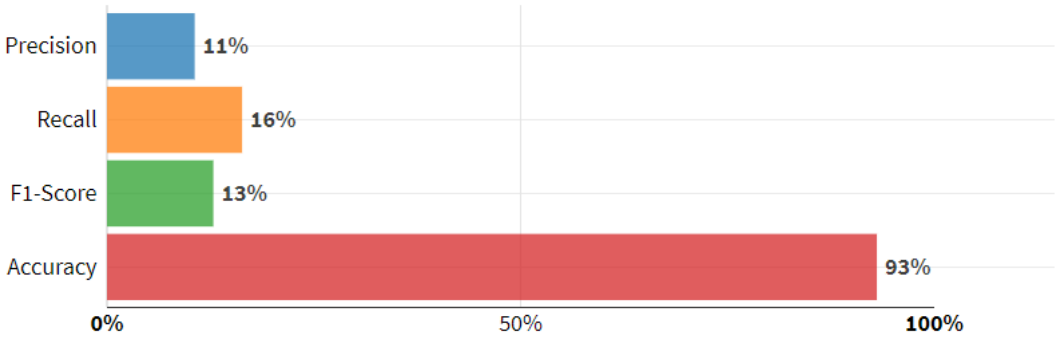

FIGURE 4.2 – les résultats de différents métriques par régression logistique

Le graphique 4.3 de densité illustre comment le modèle réussit à reconnaître (et à séparer) les classes (par exemple 1 et 0 pour une classification binaire). Il montre la répartition des classes réelles dans l'ensemble de données de test qui appartiennent réellement à la classe observée par rapport aux lignes qui n'y appartiennent pas.

Ainsi, dans ce cas, le graphique montre la distribution du taux de résiliation réel (couleur orange) par rapport à la probabilité prédite (axe des x). L'axe des y est la fonction de densité de probabilité pour l'estimation de la densité du noyau, c'est-à-dire que la densité de probabilité peut être expliquée par la probabilité par unité sur l'axe x. La couleur orange indique le taux de résiliation réelle, tandis que les couleurs bleues correspondent aux clients non résiliés. Les lignes verticales en pointillés marquent les médianes. Une situation optimale est lorsqu'une séparation claire entre les deux classes peut être identifiée et qu'il n'y a pas de chevauchement entre les deux couleurs.

En résumé, un modèle parfait sépare complètement les fonctions de densité :

- les zones colorées ne doivent pas se chevaucher,
- la fonction de densité de classe 0 doit être entièrement à gauche,
- la fonction de densité de classe 1 doit se trouver entièrement à droite,

Dans notre cas, un grand chevauchement des deux classes est identifié, ce qui n'est pas souhaité. D'autre part, on peut constater que la pente de la classe 0 (non résiliée) commence à diminuer radicalement à partir de la probabilité 0,45 (axe des x). Ainsi, en regardant les probabilités les plus élevées sur l'axe des x, la plupart des agents appartiennent en fait à la classe 1 (résilié). Même si certaine classe 0 (non résiliée) peuvent encore être trouvées à des probabilités plus élevées. Comme mentionné précédemment, le modèle parfait devrait être en mesure de séparer les modèles et n'avoir aucun chevauchement, ce qui n'est clairement pas le cas ici.

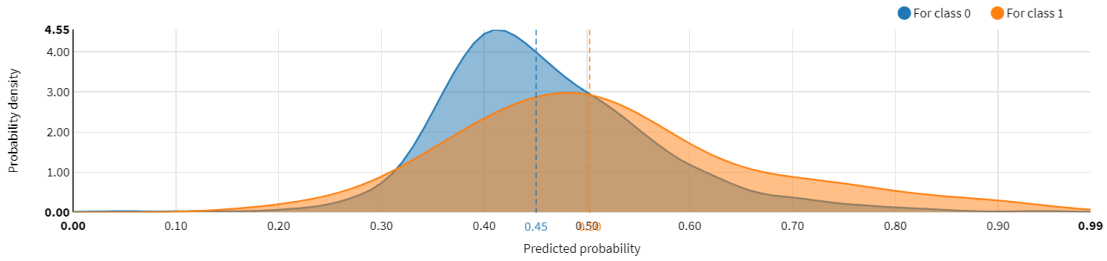

FIGURE 4.3 – Distribution de la densité par la régression logistique

4.1.3 Arbre de décision

La formation d'un arbre de décision avec ce processus automatisé peut donner lieu à de grands arbres de décision dont les sections sont très peu performantes en termes de classification. De plus, les arbres ont tendance à être sur-ajusté, ce qui signifie qu'ils s'adaptent trop étroitement aux instances de formation. Il en résulte de mauvaises performances lorsque ces arbres sont appliqués à des données non observées.

		Classe prédite		Total
		Résilié	Non-Résilié	
Classe réelle	Résilié	34	64	98
	Non-Résilié	548	2 486	3 031
Total		582	2 547	3 129

TABLE 4.3 – les résultats sur une matrice de confusion par arbre de décision

L'algorithme prédit que 582 individus ont résilié, dans les prédictions faites, il y a 548 individus mal classés, il conduit inévitablement à une précision faible, ce qui conduit également à une performance tout aussi faible, car le modèle a tendance à prédire les individus de classe majoritaire.

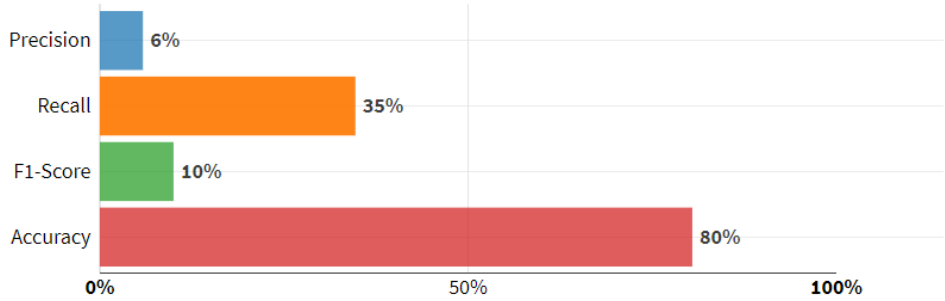

FIGURE 4.4 – les résultats de différents métriques par arbre de décision

Les probabilités estimées du modèle de l'arbre de décision sont présentées dans le graphique 4.5 ci-dessous. La fonction de densité de classe 0 se trouve au centre et à droite de l'axe x, cependant la fonction de densité de classe 1 est plus répartie le long de tout l'axe des x, ce qui n'est pas souhaité.

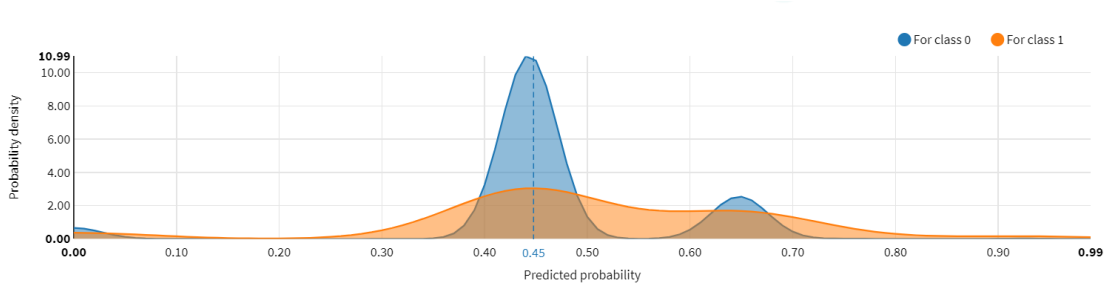

FIGURE 4.5 – Distribution de la densité par l'arbre de décision

La représentation graphique de l'arbre de décision obtenue est la suivante :

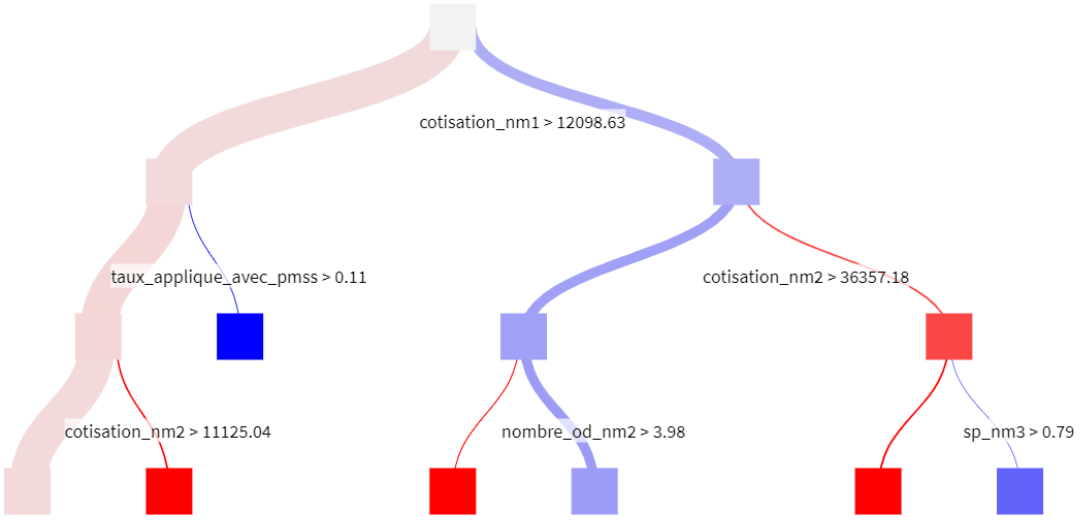

FIGURE 4.6 – Arbre de décision

Les arbres de décision recherchent l'attribut le plus important à chaque point de séparation. L'attribut le plus significatif est toujours placé comme nœud racine de l'arbre de décision donné. En regardant la figure 4.6 : L'attribut le plus important est cotisation de l'année nm1 (2020) qui est apparu dans le premier niveau de l'arbre de décision. Les autres attributs les plus importants sont la cotisation de l'année nm2(2019) et taux d'indexation appliquée pour l'année 2022.

4.1.4 Forêt aléatoire

		Classe prédite		Total
		Résilié	Non-Résilié	
Classe réelle	Résilié	53	45	98
	Non-Résilié	897	2 134	3 031
Total		950	2 179	3 129

TABLE 4.4 – les résultats sur une matrice de confusion par forêt aléatoire

Dans le tableau de confusion, nous pouvons constater que seulement 53 sur 950 résiliations ont été prédites correctement, ce qui donne une précision de 6 %.

Sur les 98 assurés résiliés observés, 53 ont été bien prédits et 45 ont été mal classé. Cela donne une sensibilité (recall) de 54 %, en termes de score de recall, ce modèle est meilleur que les autres modèles étudiés précédemment.

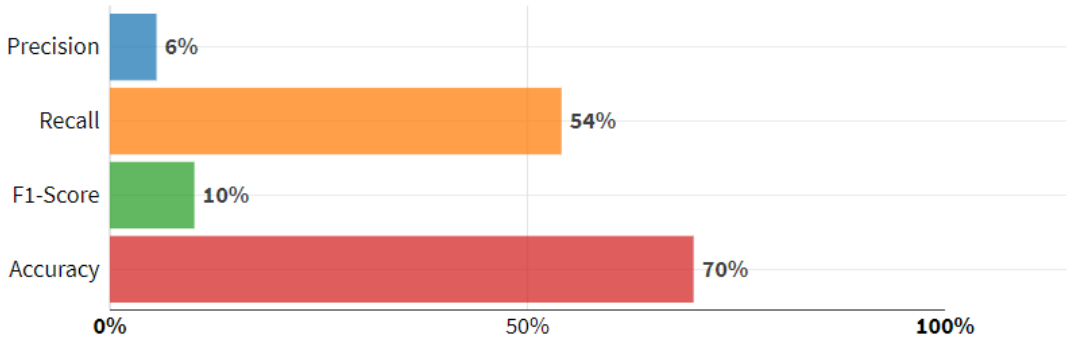

FIGURE 4.7 – les résultats de différentes métriques par forêt aléatoire

Le graphique de densité de la Figure 4.8 visualise la distribution des probabilités prédites par rapport à la classe correctement prédites. On peut en conclure que la forêt aléatoire est plus certain sur les clients prédits comme classe 0 (non-résilié). Il n'est cependant pas aussi confiant dans les prédictions de la classe 1 puisque celles-ci sont plus distribuées à gauche de l'axe des x. Cela peut être confirmé en regardant la matrice de confusion, table 4.4. Donc le modèle la forêt aléatoire semble mieux prédire les assurés actifs que les assurés résiliés. En comparant le graphique de densité pour la régression avec la forêt aléatoire, aucun des graphiques ne montre une situation optimale.

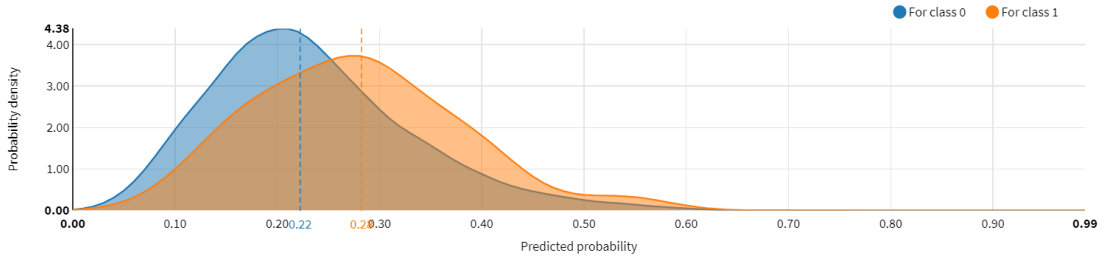

FIGURE 4.8 – Distribution de la densité par la forêt aléatoire

4.1.5 Gradient Boosting

		Classe prédite		Total
		Résilié	Non-Résilié	
Classe réelle	Résilié	24	74	98
	Non-Résilié	293	2 738	3 031
Total		317	2 812	3 129

TABLE 4.5 – les résultats sur une matrice de confusion par Gradient Boosting

En regardant la matrice de confusion, sur les 2 812 individus prédits comme non-résiliés, il y a 2 738 bonnes prédictions, il peut être conclu que le Gradient Boosting est certain sur les clients prédits comme actifs (Non-Résilié). Il n'est cependant pas aussi confiant dans les prédictions de la classe 1.

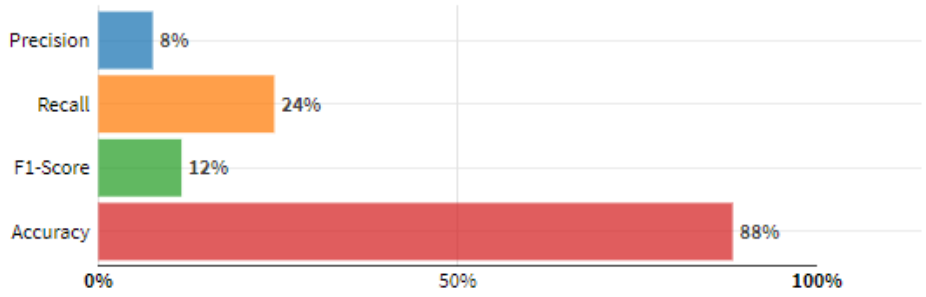

FIGURE 4.9 – les résultats de différents métriques par Gradient Boosting

Les probabilités estimées pour le modèle Gradient Boosting sont affichées dans la figure 4.10. Comme on peut le voir sur la figure, le modèle Gradient Boosting n'estime pas les probabilités de la classe 0 avec la même certitude que le modèle forêt aléatoire. Il y a toujours des signes de séparation des classes, mais comme précédemment, les chevauchements sont toujours présents.

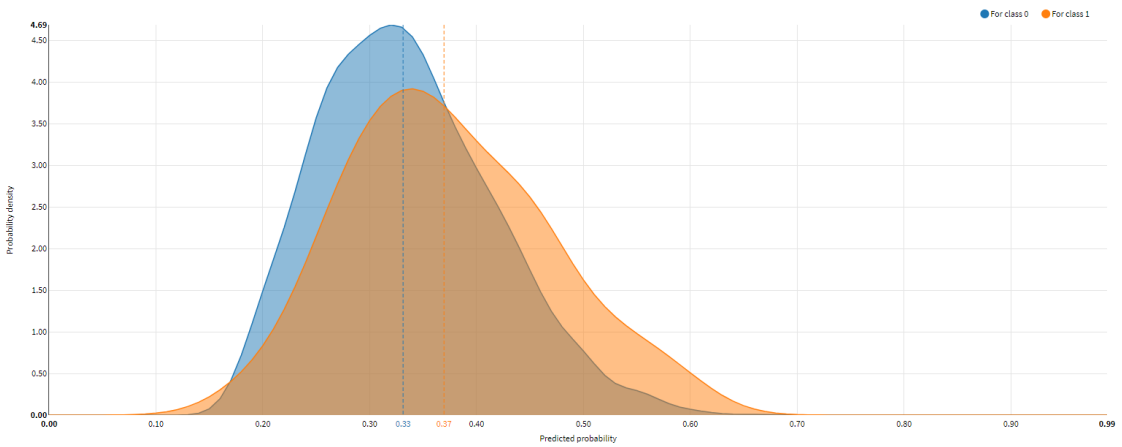

FIGURE 4.10 – Distribution de la densité par Gradient Boosting

4.1.6 Machine à vecteur support

		Classe prédite		Total
		Résilié	Non-Résilié	
Classe réel	Résilié	26	72	98
	Non-Résilié	313	2 718	3 031
Total		339	2 790	3 129

TABLE 4.6 – les résultats sur une matrice de confusion par machine à vecteur support

Le tableau 4.6 indique que 2 718 entreprises prédites comme non résiliés n'ont effectivement pas été résiliés et que 26 entreprises prédites comme résiliés ont effectivement été résiliés. L'ensemble des métriques de machine à vecteur support donne des résultats similaires à ceux du Gradient Boosting. Comme dans les autres modèles, la majorité des observations tombe dans la classe des vrais négatifs. Ce n'est pas très surprenant puisque la classe 0 est la classe majoritaire, tandis que la classe 1 est la classe minoritaire.

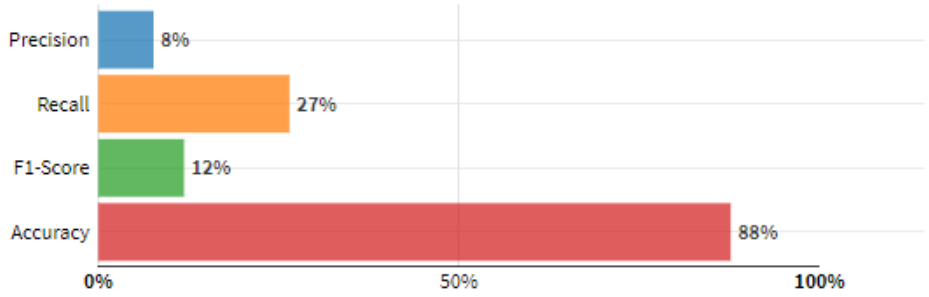

FIGURE 4.11 – les résultats de différents métriques par machine à vecteur support

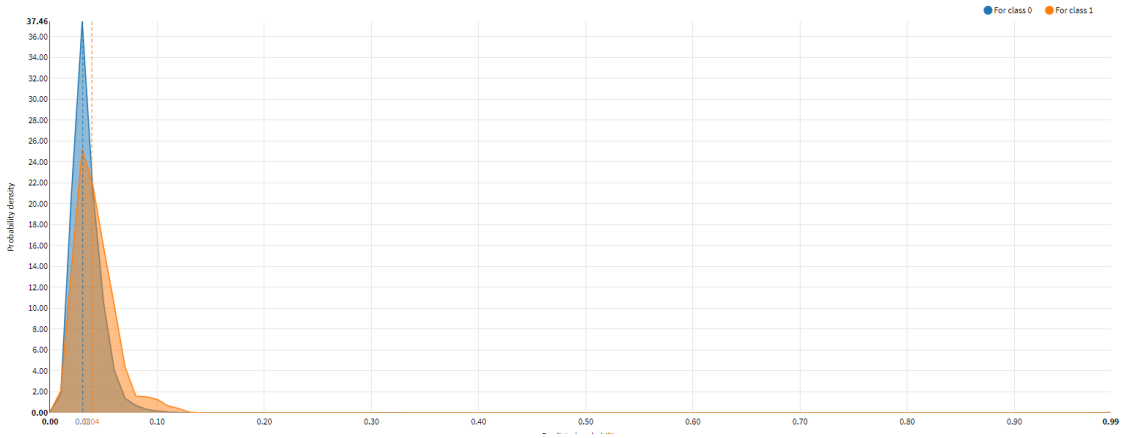

FIGURE 4.12 – la distribution des probabilités estimées par le modèle de machine à vecteur support

La distribution des probabilités estimées pour le modèle de machine à vecteur support est présentée dans la figure 4.12. Il est possible de voir les deux classes se trouvent dans la partie gauche du graphique, ça signifie que les deux classes ont des probabilités estimées proches de zéro. Il faut noter que les classes se chevauchent, ce qui signifie que la séparation n'est pas parfaite.

4.1.7 Comparaison des modèles étudiés

Les paramètres optimaux pour chaque classificateur ont été déterminés dans les sections précédentes.

La Régression logistique a une accurary supérieure à celle des autres modèles. Le modèle a également une précision et un F1-score globalement élevés que les autres modèles.

TABLE 4.7 – Récapitulation résultats des modèles étudiés

Modèle	Accurary	Précision	Recall	Score-F1	ROC AUC
Régression logistique	0.931	0.106	0.163	0.129	0.618
Arbre de décision	0.804	0.058	0.347	0.100	0.572
Forêt aléatoire	0.699	0.056	0.541	0.101	0.635
Gradient Boosting	0.883	0.076	0.245	0.116	0.620
Machine à vecteurs de support	0.877	0.077	0.265	0.119	0.639

Comme nous pouvons le constater, comparant avec les autres algorithmes, la précision et l'accuracy sont élevées pour la régression logistique, mais la capacité à identifier correctement les classes minoritaires donnée par le rappel est faible.

Dans ce mémoire spécifique, la précision fait référence au pourcentage de clients que le modèle prévoit de résilier et qui ont effectivement résilié. Par conséquent, on peut dire que la précision montre à quel point on peut-être certain des prédictions positives. Elle ne dit cependant rien sur les clients qui vont résilier mais que le modèle classe à tort comme non-résilié. Les modèles produisent une précision assez faible (voir le tableau ci-dessous).

Le recall, quant à lui, s'intéresse aux clients qui vont effectivement résilier, et combien d'entre eux ont été trouvés. Il s'agit également d'une mesure importante pour comprendre comment les modèles peuvent prédire les résils. Ici, la recall désigne l'algorithme forêt aléatoire comme meilleur modèle puisqu'il permet de cibler environ 54.1 % des résiliations. Au lieu de cela, le meilleur ROC AUC est trouvé à 63 % avec le Support

Vector Machine. Cependant, L'AUC approche les 0,6 ce qui nous amène à penser que les modèles créés ne sont pas performants.

4.2 L'interprétabilité de modèle

Étude de l'importance des variables
Le modèle de forêt aléatoire conservé au paragraphe précédent nous permet d'obtenir un score de recall le plus élevé, nous avons utilisé la forêt aléatoire pour calculer l'importance de chacune des variables, dans la prédiction des résiliations : La décroissance moyenne de l'indice de Gini. Cette mesure permet alors d'établir un classement des variables, hiérarchisées de la plus significative à la moins significative. Les résultats sont présentés dans le graphique 4.13, où les caractéristiques sont classées par ordre décroissant d'importance.

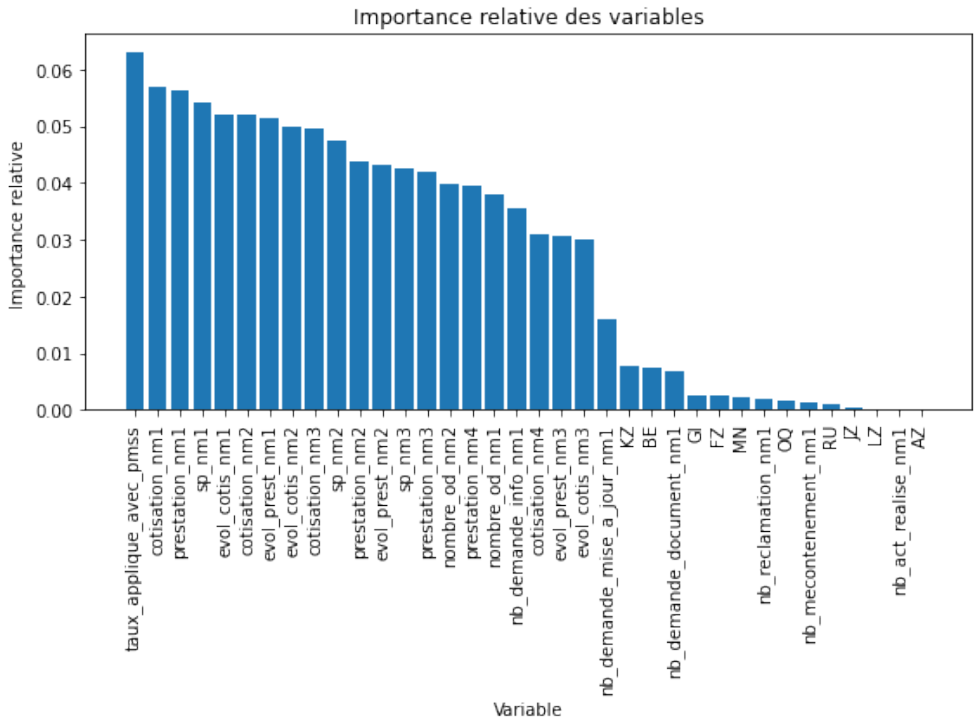

FIGURE 4.13 – Graphique de l'importance des caractéristiques classant les caractéristiques les plus importantes dans l'algorithme forêt aléatoire

On peut alors constater dans la table 4.8 que le taux d'indexation, les montant de primes et sinistres représentent les variables les plus significatives pour la résiliation, viennent ensuite les variables S sur P, l'évolution de cotisation et le niveau de couverture combiné, et enfin le nombre de contrats.

Régression logistique	Forêt aléatoire
taux d'indexation	taux d'indexation
nombre acte réalisé en 2020	cotisation en 2020
BE(a10)	prestation en 2020
LZ(a10)	S sur P en 2020
GI(a10)	évolution de cotisation en 2020

TABLE 4.8 – Variables explicatives par modèle

Nous remarquons qu'il existe des variables comme la variable cotisation et prestation dans le top 5 des variables les plus importantes dans la forêt aléatoire, mais n'ont pas été sélectionnées dans la régression logistique. Nous remarquons aussi que la variable la plus importante de modèle de forêt aléatoire (le taux d'indexation) est le même que dans la régression logistique.

Dépendance partielle

Les graphiques de dépendance partielle des quatre variables les plus explicatives sont affichés dans la figure 4.14, ils nous permettent une fois de plus de dresser le portrait type de l'individu prédit à résilier. Tout comme les variables continues en tant que telles et non par classes comme le fait la régression logistique. Les valeurs critiques des variables associés au rapprochement peuvent donc être mis en avant.

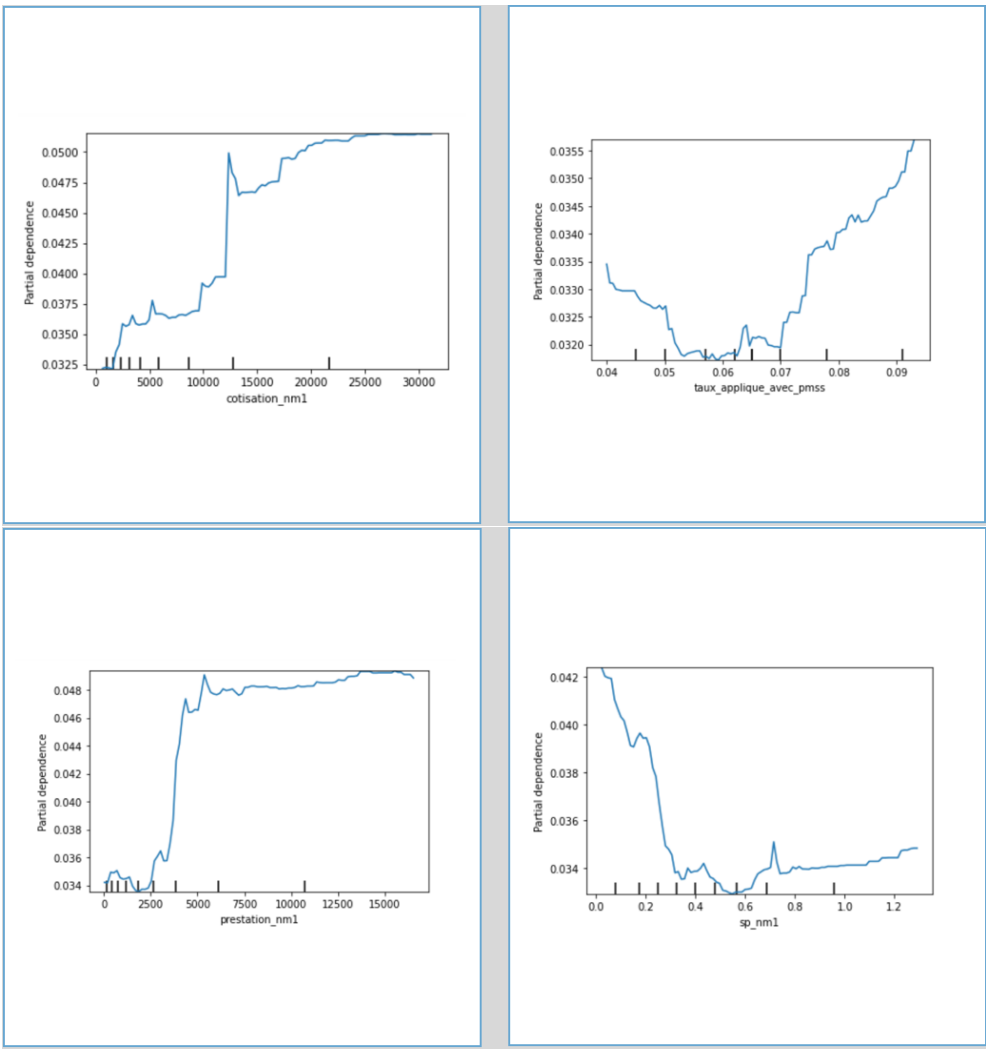

FIGURE 4.14 – Graphiques de dépendance partielle pour les variables principales

Le graphique 4.14 de PDP suggère qu'en moyenne ces quatre variables sont toutes significatives dans la prédiction de Y par le modèle de forêt aléatoire. Le graphe de la cotisation confirme le fait qu'une cotisation élevée cause une résiliation plus prononcée.

4.3 Résultats ré-échantillionage

Nous exposons les résultats de l'étude comparative des différentes méthodes de réechantillionage vu en la partie théorie. Dans cette partie, les résultats selon les différents types de ré-échantillonnage seront présentés.

En re-échantillonnant uniquement sur les données d'apprentissage, aucune des informations contenues dans les données de test n'est utilisée pour créer des observations synthétiques. L'ensemble de test aura toujours une distribution égale à la distribution de l'ensemble de données original.

Je vais me concentrer sur le rappel de la classe minoritaire (churn). Le rappel nous permet de comprendre la capacité du classificateur à identifier correctement les clients qui vont résilier.

4.3.1 Mise en œuvre du ré-échantillonnage

Quatre techniques de ré-échantillonnage ont été réalisés : sur-échantillonnage, sous-échantillonnage, SMOTE, SMOTE & Tomek-Links. Plus précisément, seul l'ensemble d'apprentissage a été ré-échantillonné, et l'ensemble de test a été conservé tel quel.

Dans cette étude, 70 % des observations ont été utilisées pour l'apprentissage, et les 30 % restants ont été utilisés pour la base de test. Pour rappel, la base d'apprentissage est constituée de 242 entreprises résiliées et de 7 017 entreprises non résiliées. La proportion de cas positifs est de 3,3 %. La distribution de résiliation d'après ré-échantillonnage sont les suivants :

TABLE 4.9 – Répartition des effectifs des bases ré-échantillonnées

	Non-Résilié (= 0)	Résilié (= 1)	Total
Base déséquilibrée	7 017 (96.7 %)	242 (3.3 %)	7 259
Sur-échantillonnage	7 017 (50 %)	7 017 (50 %)	14 034
Sous-échantillonnage	242 (50 %)	242 (50 %)	484
SMOTE	7 017 (50 %)	7 017 (50 %)	14 034
SMOTE &TomekLinks	6 995 (50 %)	6 995 (50 %)	13 990

La base est ré-échantillonnée de telle sorte que les nombres de l'entreprise non résiliée et les nombres de l'entreprise résiliée sont équivalentes.

Le jeu de données ré-échantillonné par sur-échantillonnage est composé de 7 017 cas positifs et de 7 017 cas négatifs. Ainsi, la proportion de cas positifs est de 50 %. Le déséquilibre des classes a été parfaitement corrigé. Le nombre d'observations dans les données de l'apprentissage sous-échantillonné est de 484, et 242 d'entre elles sont des cas positifs. L'application de SMOTE augmente le nombre de cas positifs. Le jeu de données résultant contient 7 017 cas positifs et 7 017 cas négatifs. Ainsi, l'ensemble de données est parfaitement équilibré. Le jeu de données ré-échantillonné par SMOTE & TomekLinks est composé de 6 995 cas positifs

et de 6 995 cas négatifs. Le déséquilibre de classe a été corrigé, mais l'observation totale est plus petite que le jeu de données rééquilibré par SMOTE car SMOTE & Tomek-Links génère des cas positifs synthétiques et supprime certains cas négatifs.

4.3.2 Comparaison des méthodes de ré-échantillonnage

Un ensemble de données déséquilibré peut poser certains problèmes lorsqu'un modèle est évalué sur l'accuracy. Par exemple, si la population est composée d'un ratio de classes positives et négatives de 1 : 9 et que le modèle classe chaque client comme négatif, l'accuracy est de 90 %. Donc il est plus intéressant de voir F1 scores.

$$F1 - score = \frac{2 \times \text{ recall } \times \text{ precision}}{\text{recall+precision}}$$

Le tableau suivant montre les scores de F1 calculés à partir d'ensembles de base de données de test pour chaque modèle.

TABLE 4.10 – Table : F1 score (sur l'échantillon de test) des divers modèles d'apprentissage combinés avec les techniques de rééchantillonnage

	F1 score				
	Base originale	Sur-échantillonnage aléatoire	Sous-échantillonnage aléatoire	Smote	Smote & TomekLinks
Régression logistique	0.129	0.096	0.065	0.016	0.018
Arbre de décision	0.096	0.010	0.064	0.006	0.060
Forêt aléatoire	0.101	0	0.065	0	0
Gradient Boosting	0.116	0.068	0.065	0.036	0.017
Machine à vecteurs de support	0.119	0.060	0.064	0.064	0

Tout d'abord, comparons les méthodes d'ensemble, de façon générale, 4 modèles de re-échantillonnage restent faibles, même moins bien que l'ensemble de données initial déséquilibrées.

Ensuite, nous évaluerons combien de résiliation sont correctement prédits. Ce taux est calculé avec l'équation suivante, et représente le taux de vrais positifs.

$$Recall(Rappel) = \frac{TP}{TP + FN}$$

Pour valider les modèles et leurs performances, il est intéressant de regarder leurs performances en termes de prédiction. Afin d'évaluer les performances des algorithmes en considérant le rappel.

Le tableau 4.11 montre les scores de recall calculés à partir d'ensembles de base de données de test pour chaque modèle. Tout d'abord, comparons les méthodes d'ensemble, on arrive à une conclusion que le sous-échantillonnage a des mesures de performance supérieures aux autres techniques de ré-échantillonnage. De façon générale, 3 modèles de sur-échantillonnage restent faibles, même moins bien que l'ensemble de données déséquilibrées. Il s'agit du sur-apprentissage aléatoire, smote et SMOTE & Tomek.

D'après le score de recall, pour le jeu de données déséquilibré, la forêt aléatoire présente le meilleur recall (0,541). Pour les jeux de données rééchantillonnées, la forêt aléatoire a obtenu le meilleur score 0.923 dans sous-échantillonnage aléatoire, et l'arbre de décision a obtenu le meilleur score dans SMOTE.

TABLE 4.11 – Récapitulation résultats avec sous-échantillonnage

	Recall				
	Base originale	Sur-échantillonnage aléatoire	Sous-échantillonnage aléatoire	Smote	Smote & TomekLinks
Régression logistique	0.163	0.519	0.910	0.038	0.01
Arbre de décision	0.347	0.010	0.903	0.558	0.077
Forêt aléatoire	0.541	0	0.923	0	0
Gradient Boosting	0.245	0.202	0.913	0.019	0.009
Machine à vecteurs de support	0.265	0.250	0.817	0.250	0.26

Parmi les algorithmes de classification, la régression logistique a un meilleur score de recall pour l'ensemble de données sur-échantillonnées aléatoirement. Cependant, l'arbre de décision a obtenu le meilleur recall pour les deux ensembles de données ré-échantillonnés. Pour l'ensemble de données déséquilibrées, le forêt aléatoire présente le score le plus bas parmi les méthodes d'ensemble de sur-échantillonnage, mais a montré des scores plus élevés pour les ensembles de données originaux et les ensembles de données sous-échantillonnées. Comparé aux autres algorithmes de classification, Grandient Boosting et Machine à vecteurs de support ont montré les scores de recall les plus faibles pour chaque ensemble de données.

Ces résultats montrent que travailler sur la manière de sous-échantillonner améliore la prédiction de classe résilié. Donc la méthode de sous-échantillonnage aléatoire est à privilégier.

4.3.3 Comparaison des modèles avec sous-échantillonnage

Dans cette partie, afin d'avoir une idée de la qualité des modèles avec méthodes de sous-échantillonnage, les matrices de confusion sont calculés sur différents modèles.

Les résultats pour les cinq algorithmes sont présentés dans les tableaux, à gauche, nous avons les résultats obtenus à partir de base déséquilibrée sans ré-échantillonnage, à droite, les matrices de confusion montrent des résultats sur la base ré-échantillonnée.

TP(vrai positif) se réfère à la minorité (Résilié) des valeurs correctement classées ; la méthode de ré-échantillonnage augmente le nombre de ces valeurs, ce qui porte la valeur de la sensibilité à environ de 90 %. En comparant le nombre de faux positifs, on constate que le nombre de TN (vrai négatif) dans le sous-échantillonnage a beaucoup diminué que dans les autres méthodes. TN signifie la classification réelle des instances "Non-Résilié", où la valeur "Non-Résilié" fait référence à la classe majoritaire, ce qui signifie que le sous-échantillonnage peut renforcer les classes minoritaires en améliorant leurs échantillons. Cependant, le sous-échantillonnage se trompe dans la prédiction de la classe majoritaire.

Régression logistique			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	16	82	98
Non-Résilié	135	2 896	3 031
Total	151	2 978	3 129

TABLE 4.12 – les résultats sur une matrice de confusion sans ré-échantillonnage.

Régression logistique			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	89	9	98
Non-Résilié	2 308	723	3 031
Total	2 397	732	3 129

TABLE 4.15 – les résultats sur une matrice de confusion avec sous-échantillonnage.

Arbre de décision			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	34	64	98
Non-Résilié	548	2 486	3 031
Total	582	2 547	3 129

TABLE 4.13 – les résultats sur une matrice de confusion sans re-échantillonnage.

Arbre de décision			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	94	4	98
Non-Résilié	2 463	568	3 031
Total	2 557	572	3 129

TABLE 4.16 – les résultats sur une matrice de confusion avec sous-échantillonnage.

Forêt aléatoire			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	53	54	98
Non-Résilié	897	2 134	3 031
Total	950	2 179	3 129

TABLE 4.14 – les résultats sur une matrice de confusion sans ré-échantillonnage.

Forêt aléatoire			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	91	7	98
Non-Résilié	2 390	641	3 031
Total	2 397	648	3 129

TABLE 4.17 – les résultats sur une matrice de confusion avec sous-échantillonnage.

Machine à vecteurs de support			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	26	72	98
Non-Résilié	313	2 718	3 031
Total	339	2 790	3 129

TABLE 4.18 – les résultats sur une matrice de confusion sans ré-échantillonnage.

Machine à vecteurs de support			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	90	8	98
Non-Résilié	2 324	707	3 031
Total	2 414	715	3 129

TABLE 4.20 – les résultats sur une matrice de confusion avec sous-échantillonnage.

Gradient Boosting			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	24	74	98
Non-Résilié	293	2 738	3 031
Total	317	2 812	3 129

TABLE 4.19 – les résultats sur une matrice de confusion sans ré-échantillonnage.

Gradient Boosting			
	Classe prédite		
Classe réelle	Résilié	Non-Résilié	Total
Résilié	92	6	98
Non-Résilié	2 328	703	3 031
Total	2 420	709	3 129

TABLE 4.21 – les résultats sur une matrice de confusion avec sous-échantillonnage.

Prenons comme exemple les matrices de confusion de Machine à vecteurs de support. Après le ré-échantillonnage, la classification du nouvel ensemble de données a considérablement augmenté le nombre de vrais positives, de 26 à 90, le nombre de faux positives a également monté de 313 à 2324. Les modèles ont une grande amélioration dans la prédiction de classe résiliée avec la méthode sous-échantillonnage aléatoire, car les algorithmes ont pu détecter des classes minoritaires. Sur les 98 assurés résiliés observés, environ 90 ont été bien prédits et très peu ont été mal classé. Cela donne une **sensibilité (recall)** de $\frac{90}{98} \approx 92\%$.

Supposons que nous puissions garder tous les clients après avoir prévu qu'ils vont résilier.
Coût sans ré-échantillonnage = perte de 72 clients + gaspillage d'efforts marketing et d'argent sur 313 clients parce que nous pensions les perdre.
Coût du sous-échantillonnage = perte de 8 clients + gaspillage sur 2 324 clients.

4.4 Limites

Les limites de la base de données auraient pu avoir un effet important sur le résultat de la performance du modèle, car l'accent était mis sur les informations historiques et non sur les informations caractéristiques des clients. Avec des informations plus spécifiques au client, la performance pourrait augmenter. Lorsqu'il s'agit d'une question très spécifique au client, telle que comme la question de la résiliation, le gain relatif lié à l'utilisation de plus de variables liées au client serait très probablement bénéfique aux performances du modèle.

Conclusion du chapitre
Cette partie a permis de tester plusieurs algorithmes sur les données : GLM, arbre de décision, forêt aléatoire,

GBoost, SVM. Ils montrent tous cinq des résultats ne sont pas perfectibles, même si forêt aléatoire semble être légèrement meilleur que les autres modèles pour la détection des individus prêts à résilier.

Certaines caractéristiques sont plus importantes que d'autres. Dans le cas de forêt aléatoire, le taux d'indexation est la variable la plus significative. Cela était attendu en raison du fait que les clients allaient résilier les contrats s'ils ont reçu un taux d'augmentation élevée.

Pour comparer les résultats, nous appliquons les cinq modèles sur différentes bases ré-échantillonnées et nous calculons l'indicateur recall. Les meilleurs résultats de score de recall s'observent pour le modèle sous-apprentissage aléatoire. Le sous-échantillonnage répond le mieux à notre problème dans ce cas, car elles élargissent la voix de la classe minoritaire.

Chapitre 5

Mesure de l'élasticité

Pour chaque assuré, nous avons donc estimé la probabilité de résiliation et son élasticité. Il est de ce fait possible d'étudier les résultats globalement, ou par segment de population suivant la ou les variables que l'on juge intéressantes.
Deux caractéristiques importantes à connaître sur un assuré sont son risque de résiliation et sa sensibilité à l'augmentation de prix. L'estimation de ces deux caractéristiques et la catégorisation des assurées en quatre quadrants définis par ces deux mesures peuvent aider à déterminer la meilleure tarification à adopter avec chaque client.

5.1 Calcul de l'élasticité de la probabilité de résiliation à la hausse de taux d'indexation

L'élasticité de la probabilité de résiliation au taux d'indexation correspond à la variation relative de cette probabilité lorsque le taux d'indexation varie d'une quantité infinitésimale. Ceci permet de connaître la réaction des assurés face à une variation de taux d'indexation, en termes de résiliations.
Pour connaître l'élasticité du modèle. On peut alors écrire :

$$\hat{\mathscr{E}}(t) = \frac{\partial \hat{\tau}(t)}{\hat{\tau}(t)} / \frac{\partial t}{t} = \frac{\partial \hat{\tau}(t)}{\partial t} \times \frac{t}{\hat{\tau}(t)}$$

Où $\hat{\tau}(t)$ est le taux de résiliation estimé en fonction du taux d'indexation, t est le taux d'indexation.

Dans la partie théorique, nous avons introduit la fonction logit, qui fournit une formule directe de la probabilité de résiliation, qui est spécifiée par des coefficients.
En utilisant $P(Résil)$ et $P(Non-Résil)$, nous pouvons spécifier odds comme :

$$Odds(Résil) = \frac{P(Résil)}{P(Non-Résil)} = \frac{P(Résil)}{1-P(Résil)}$$

Les odds de résiliation sont comprises entre 0 et ∞. Ainsi,

$$logit(Odds(Résil)) = log(\frac{P(Résil)}{1-P(Résil)})$$

Comme la valeur de logit est entre $-\infty$ et ∞, nous pouvons les relier à n'importe quelle variable indépendante et interpréter les effets des variables, sauf que l'effet des variables serait sur logit de la variable cible. Donc nous pouvons écrire l'équation réliant logit de la probabilité de résiliation avec les variables taux d'indexation :

$$logit(Odds(Résil)) = \alpha + \beta t$$

nous définissons l'estimateur du taux de résiliation attendu s'écrit :

$$\hat{\tau}(t) = \frac{1}{1+e^{-(\alpha+\beta t)}}$$

Avec :

α : le coefficient associé à la constante, qui correspond donc aux caractéristiques de l'assuré et non au taux d'indexation,]

β : le coefficient associé au niveau de taux d'indexation t, un coefficient lié aux caractéristiques de l'assuré et au taux d'indexation.

Dans notre base de données nous disposons donc des variables qui nous a permis d'estimer les coefficients α et β grâce à une régression logistique avec fonction de lien logit.

Nous calculons la dérivée de la fonction :

$$\frac{\partial \hat{\tau}(t)}{\partial t} = \frac{\beta e^{-(\alpha+\beta t)}}{\left(1+e^{-(\alpha+\beta t)}\right)^2}$$

Il est aussi possible, avec la forme de la probabilité de résiliation de calculer une sensibilité de chaque assuré à son niveau de taux d'indexation. Pour cela, il suffit de dériver la fonction qui permet de calculer cette probabilité en fonction du niveau de taux d'indexation.

$$\begin{aligned}
\hat{\mathscr{E}}(t) &= \frac{\partial \hat{\tau}(t)}{\hat{\tau}(t)} / \frac{\partial t}{t} \\
&= \frac{\partial \hat{\tau}(t)}{\partial t} \times \frac{t}{\hat{\tau}(t)} \\
&= \frac{\beta e^{-(\alpha+\beta t)}}{\left(1+e^{-(\alpha+\beta t)}\right)^2} \times \frac{t}{\frac{1}{1+e^{-(\alpha+\beta t)}}} \\
&= \frac{\beta t e^{-(\alpha+\beta t)}}{1+e^{-(\alpha+\beta t)}} \\
&= (1-\hat{\tau}(t))\beta t
\end{aligned}$$

Donc nous obtenons une formule explicite par assuré de l'élasticité de la probabilité de résiliation

$$\hat{\varepsilon}(t) = \beta t(1-\hat{\tau}(t))$$

Nous faisons l'hypothèse que le coefficient β est positif.
Une élasticité presque nulle indique donc que l'assuré n'est absolument pas sensible à la variation de prix. Ainsi, si son tarif varie, sa probabilité de résiliation n'en sera pas affectée. A l'inverse, si l'élasticité est élevée en valeur absolue, l'assuré est alors très sensible à la variation du tarif.

5.2 Résultats

Pour chaque assuré, nous avons donc estimé la probabilité de résiliation et son élasticité.

En 2021
Si on revient sur les résultats de nos données, on obtient les résultats suivants :

Résumé des résultats pour l'année 2021	
α	-3.8739
β	7.5570
P value	0

Les p-values associées au test de l'hypothèse nulle, considérant que tous les coefficients associés aux variables explicatives sont nuls.

Dans le cas présent la probabilité de résiliation : $\hat{\tau}(t) = \frac{1}{1+e^{-(3.8739+7.5570t)}}$
l'élasticité : $\hat{\varepsilon}(t) = \hat{\tau}(t) \times 7.5570 \times t$

La figure montre les régressions de la résiliation estimée en utilisant les taux d'indexation calculées à partir des données de 2021.

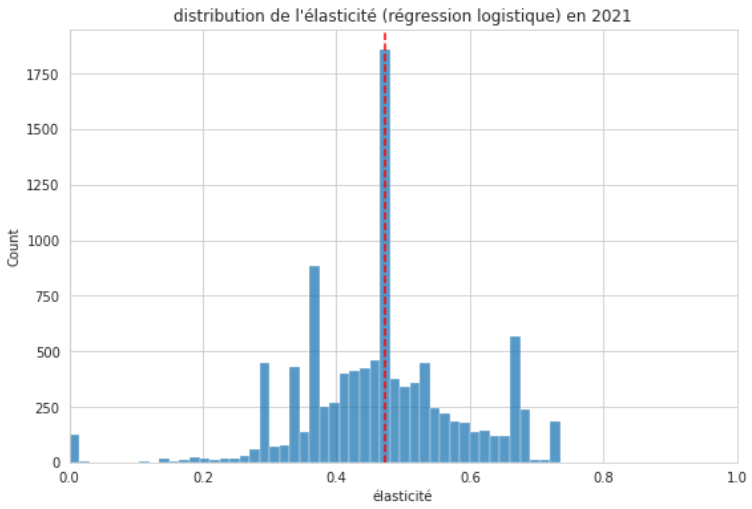

FIGURE 5.1 – La distribution de l'élasticité

Trois remarques peuvent être faites :

- L'élasticité est positive : une augmentation du taux d'indexation entraînerait donc une augmentation des résiliations.
- L'élasticité estimée est dans la fourchette de 0 à 0,8.
- L'élasticité est très proche de 0.5 au niveau de la médiane : cela veut dire qu'une augmentation de 1 % du taux d'indexation va augmenter la probabilité de résiliation de 0.5 %

.

En 2020
On reprend le même calcul que précédemment avec les taux d'indexation de 2020.

Résumé des résultats pour l'année 2020	
α	-2.8549
β	0.7770
P value	0

Le coefficient β de taux d'indexation est positive et statistiquement significative.

Dans le cas présent la probabilité de résiliation : $\hat{\tau}(t) = \frac{1}{1+e^{-(-2.8549+0.7770t)}}$
l'élasticité : $\hat{\varepsilon}(t) = \hat{\tau}(t) \times 0.7770 \times t$

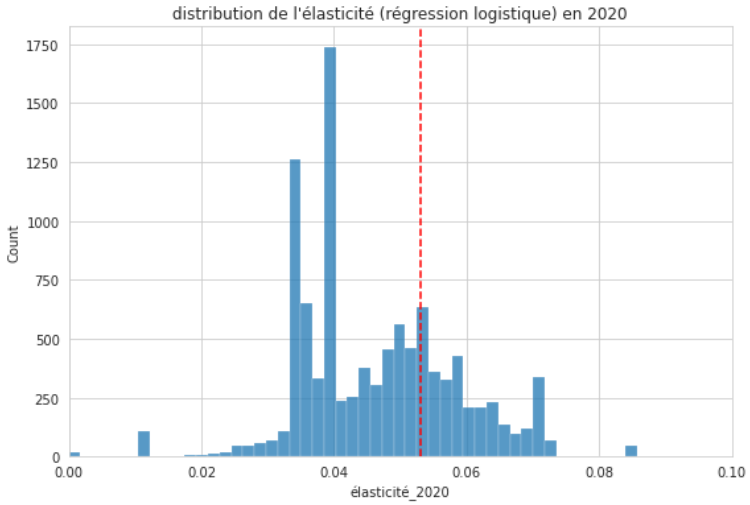

FIGURE 5.2 – La distribution de l'élasticité en 2020

Le graphique 5.2 montre que nous obtenons des élasticités positives, ce qui est cohérent. L'élasticité-prix estimée est d'environ 0,05. Cependant, les valeurs estimées des élasticités-prix en 2020 varient peu ; l'élasticité-prix varie généralement entre 0,03 et 0,08.

En 2019
La régression logistique permet d'obtenir les coefficients et nécessaires au calcul de l'élasticité. Les valeurs sont présentés dans la table dessous :

Résumé des résultats pour l'année 2019	
α	-3.0006
β	3.5904
P value	0

Dans le cas présent la probabilité de résiliation : $\hat{\tau}(t) = \frac{1}{1+e^{-(-3.0006+3.5904t)}}$
l'élasticité : $\hat{\varepsilon}(t) = \hat{\tau}(t) \times 3.5904 \times t$

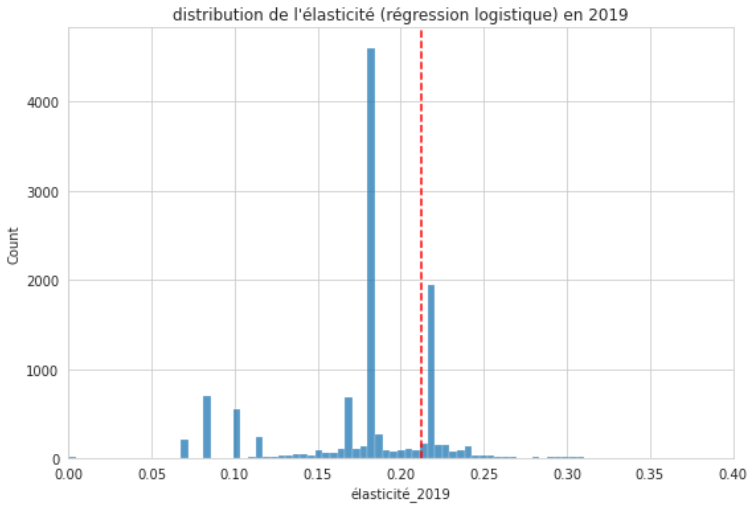

FIGURE 5.3 – La distribution de l'élasticité

L'élasticité moyenne observée en 2019 est de 0,21 d'après la figure 5.3. Cela signifie que si le niveau de taux d'indexation augmente de 1 % l'année suivante, supposons le taux d'indexation de concurrence reste inchangé, le taux de résiliation venant de l'assuré augmentera de 0.21 %.

Comparaison de la probabilité de résiliation et l'élasticité estimée de différentes années
Pour le modèle logistique de l'élasticité au taux d'indexation, nous avons considéré les taux d'indexation entre 2019 et 2021. Le graphique 5.4 montre l'évolution de probabilité de résiliation estimée de ces trois années par rapport au taux d'indexation.

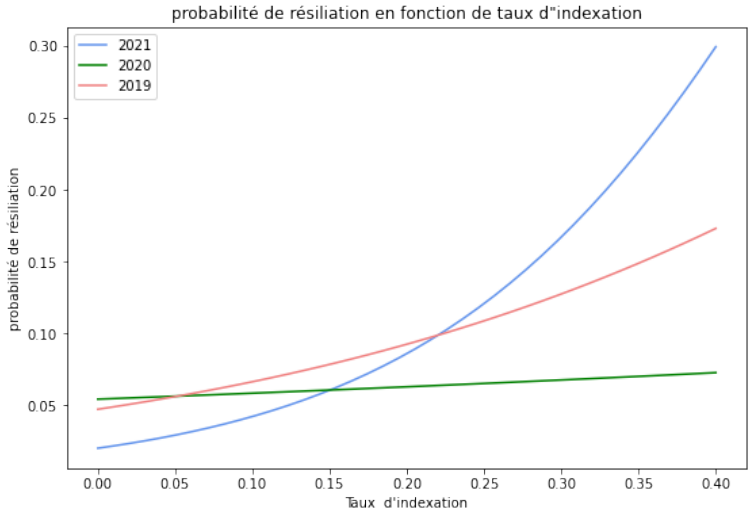

FIGURE 5.4 – Résiliation en fonction de taux d'indexation

Les évolutions des trois courbes sont toutes monotones croissantes. De plus, le graphique nous montre que plus le taux d'indexation est élevé et plus la probabilité de résiliation est forte. La probabilité de résiliation est donc sensible à cette variable.

La distribution de l'élasticité pour l'année 2019, 2020, et 2021 est décrite dans les graphiques en Figure 5.5.

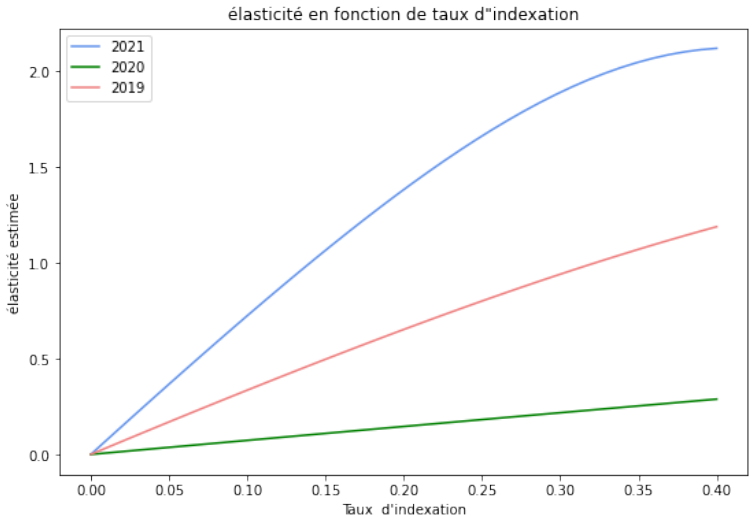

FIGURE 5.5 – Elasticité en fonction de taux d'indexation

Pour un même niveau de taux d'indexation, l'élasticité en 2021 est plus forte que celle de 2019 et 2020. En plus, l'élasticité par rapport au taux d'indexation est généralement positive et statistiquement significative, cela indique qu'une augmentation du taux d'indexation aurait donc un effet sur les résiliations. De plus, sa valeur varie de 0 à 0.2.

La figure 5.6 nous montre que la distribution de l'élasticité de 2019 à 2021.

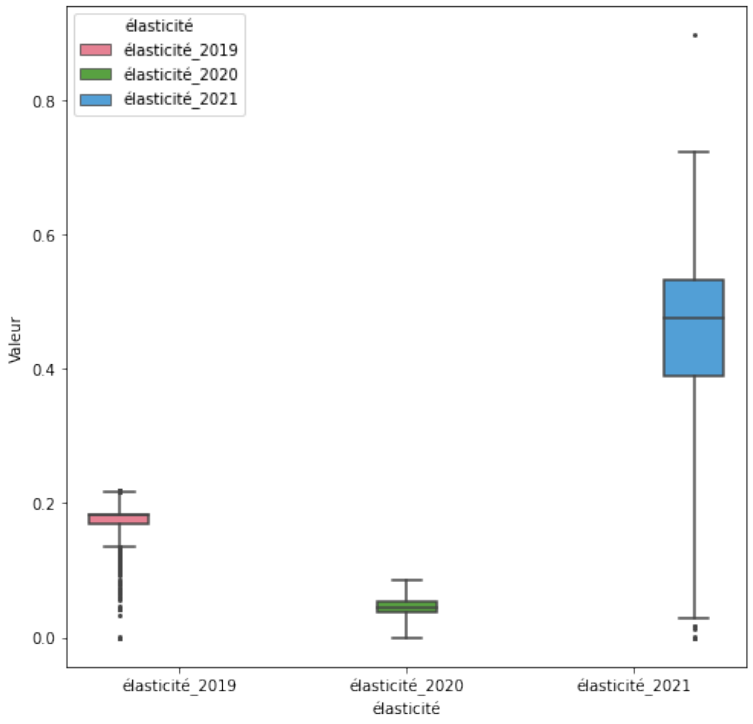

FIGURE 5.6 – Distribution de l'élasticité de 2019 à 2021

Les élasticités ont tendance à être plus élevées en 2021 pour l'ensemble des entreprises. Nous observons également que l'élasticité en global est beaucoup plus faible pour l'année 2020, ce n'est pas en contradiction avec la réalité. Parce que lors du confinement, plus de quatre sociétés sur cinq estiment avoir subi une baisse de leur activité. Dans ce contexte, les prestations ont diminué en 2020, mais les clients n'ont pas reçu une réduction de tarif, il est donc raisonnable de supposer que le client va chercher un meilleur tarif ailleurs. Cela peut par ailleurs être vérifié par les résultats obtenus en 2021, on constate une forte augmentation de l'élasticité en 2021.

En résumé, Ces résultats indiquent que les entreprises assurées font attention au changement de taux d'indexation auxquelles ils sont confrontés et à leur propre risque d'avoir besoin de prestations. Cependant, bien que les gens réagissent aux primes et aux prestations attendues, leur réaction est d'une ampleur relativement modeste.

Conclusion

L'objectif de ce mémoire était d'identifier les profils de clients qui ont une probabilité de résiliation plus élevée dans le périmètre de santé collective obligatoire.

On peut conclure que tous les modèles ont généré de résultats insatisfaisants. Cependant, le modèle forêt aléatoire a généré le score recall le plus élevé, il est donc considéré comme le modèle le plus fiable dans ce mémoire. En analysant l'importance de variable, la caractéristique la plus importante est le taux d'indexation, ce qui n'est pas surprenant car cela signifie que si un client n'est pas content de l'augmentation de taux d'indexation, qu'il va résilier son contrat. En regardant les diagrammes de densité, il est clair qu'aucun modèle ne peuvent séparer clairement les deux classes. Cela doit être pris en considération. Néanmoins, l'utilisation d'un modèle d'apprentissage automatique peut toujours aider l'entreprise à conserver les clients actuels en mettant en place des stratégies marketing avant qu'il ne soit trop tard.

En appliquant les quatre techniques de ré-échantillonnage sur la base déséquilibrée, d'après le résultat, il est également visible que la méthode de sous-échantillonnage aléatoire a un score de recall relativement élevée. Ceci est important d'un point de vue commercial, car ces informations permettent de conclure que des actions à prendre pour les clients dont on prédit qu'ils vont résilier.

L'élasticité de la probabilité de résiliation au taux d'indexation permet de connaître la réaction des assurés face à une variation de taux d'indexation. Dans la dernière partie de modélisation, on vient de voir que l'élasticité par rapport au taux d'indexation est généralement positive et statistiquement significative. De plus, pour un même niveau de taux d'indexation, l'élasticité en 2021 est plus forte que celle de 2019 et 2020.

Limites
Le secteur de l'assurance santé en général étant très dépendant des relations avec les clients, il est difficile de prévoir leurs résiliations. Normalement les relations entre les assureurs et les assurés ont un impact incontournable sur le taux de résiliation, mais il n'existe pas de moyen direct et exhaustif de les mesurer. Les variable appelée nombre de demandes de mise à jour, nombre de réclamations, et nombre de mécontentements etc sont une tentative d'inclure cet aspect dans le modèle. Cependant, il s'agit uniquement du point de vue de l'assureur sur la relation et non de la perspective de l'assuré. D'ailleurs la concurrence, c'est un élément qui n'est pas pris en compte dans la prédiction. En général, on peut conclure que ces aspects sont difficiles à inclure dans un modèle statistique.

Il est essentiel d'identifier les clients qui présentent un risque élevé de résiliation, pour avoir une chance de les conserver. Pour que la prédiction du taux de résiliation apporte de la valeur à l'entreprise, on suppose que des mesures spécifiques peuvent être fait pour empêcher les clients de partir. Cela peut être par exemple de mener des stratégies de marketing, de proposer des offres spécifiques aux clients, ou quelque chose d'aussi simple que de les appeler ou de leur envoyer un e-mail pour discuter à nouveau le niveau de garantie. Néanmoins, il existe un risque que les clients se détournent de l'assureur malgré l'assistance fournie pour les garder. La question de savoir comment le support peut être utilisé pour prévenir la résiliation est une question à considérer après l'étape de prédiction.

Il est possible qu'un meilleur résultat ait pu être généré si plus de données avaient été accessibles pour ce projet. Comme l'ensemble de données utilisé pour les prédictions est fortement déséquilibré, donc on aurait dû augmenter la période de l'observation. D'ailleurs, il aurait été préférable de disposer de plus d'informations sur les clients qui résilient.

Bibliographie

[1] Aba DIOP. 2012. Inférence statistique dans le modèle de régression logistique avec fraction immune. THÈSE, L'UNIVERSITÉ GASTON BERGER ET DE L'UNIVERSITÉ DE LA ROCHELLE. 7-11

[2] Ozlem Karatekin. 2020. Binary Tarification et mesure de l'antiselection en assurance sante collective. Mémoire d'actuariat, Universitaire d'Actuaire de Strasbourg. 18.

[3] Jennifer Karlberg Maja Axén. 2020. Binary Classification for Predicting Customer Churn. Master Thesis, Umeå University. 11.

[4] Arnaud PELTIER. 2019. Modélisation des taux de démission pour des contrats collectifs de Prévoyance-Santé sur-mesure. Mémoire d'actuariat, l'Institut du Risk Management. 38.

[5] Clara ADICEOM. 2019. Optimisation de la stratégie de majoration des primes de contrats d'assurance habitation au terme. Mémoire d'actuariat, Essec. 67.

[6] Ivan Herboch. 2016. Predictive Analytics en actuariat : application à la modélisation de la résiliation non-vie. Mémoire d'actuariat, Université Paris Dauphine. 94.

[7] Jérôme Sourisseau. 2019. Modélisation du taux de transformation et de l'élasticité prix en affaire nouvelle pour l'assurance automobile. Mémoire d'actuariat, Centre d'Etudes Actuarielles. 38.

[8] Audrey Peyriller. 2019. Prédiction des résiliations en santé individuelle. Mémoire d'actuariat, l'Institut du Risk Management. 5.

[9] Sophie GOURLIER. 2014. Analyse de la rentabilité d'un produit en santé individuelle. Mémoire d'actuariat, Université Paris Dauphine. 23-25.

[10] Yufei LUO. 2015. Amélioration de la modélisation de sinistres graves 'à l'aide d'une approche d'apprentissage. Mémoire d'actuariat, ISFA. 29.

[11] Thomas BOUCHÉ. 2014. Modèle de propension des assurés par rapport aux risques de sinistres corporels graves en assurance automobile. Mémoire d'actuariat, EURIA. 60.

[12] ORNELLIA DJOFFON. 2016. Modélisation de la survenance d'un sinistre dans le cas d'une asymétrie des classes et utilisation dans le cadre d'un modèle interne partiel. Mémoire d'actuariat, ISFA. 52-55.

[13] Lena Schütte. 2022. A Churn Model for Swiss Mandatory Health Insurance. Master Thesis, ETH Zurich. 47.

[14] Chantine Huigevoort. 2015. Customer churn prediction for an insurance company. Master Thesis, Eindhoven University of Technology. 40-43.

[15] Ossama Chihi. 2011. Sensibilité du taux de résiliation au prix en assurance MRH occupant et simulation du portefeuille. Mémoire d'actuariat, ISUP, Sorbonne université. 139.

[16] Dutang, C. 2011. Regression models of price elasticity in non-life insurance, Master's thesis, ISFA. Mémoire confidentiel - AXA Group Risk Management.

[17] Léonard Fontaine. Modélisation de la valeur contrat par la refonte du modèle de résiliation. Mémoire d'actuariat, ISFA, Université de Lyon 1, 2011. 139

Annexe

Mécanisme remboursement Sécurité Sociale

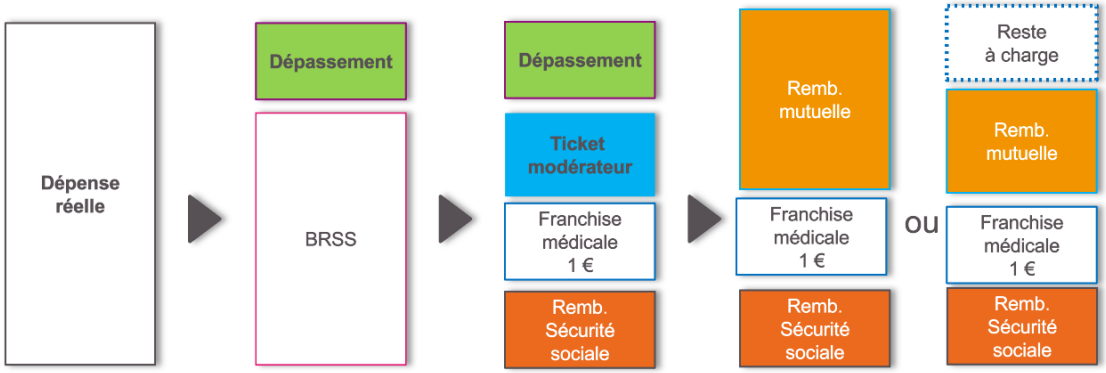

FIGURE 5.7 – Mécanisme remboursement Sécurité Sociale

Loi Evin et Accord National Interprofessionnel (ANI)

	Loi Evin	ANI
Couverture	Santé	Santé et prévoyance
Assuré(s)	Le salarié	Le salarié et ses ayants droit
Durée	Dispositif viager	Durée temporaire de 12 mois maximum
Point de départ	Le lendemain de la date de la demande du salarié	Date de la cession du contrat de travail

TABLE 5.1 – Loi Evin et Accord National Interprofessionnel (ANI)

Le 100 % santé

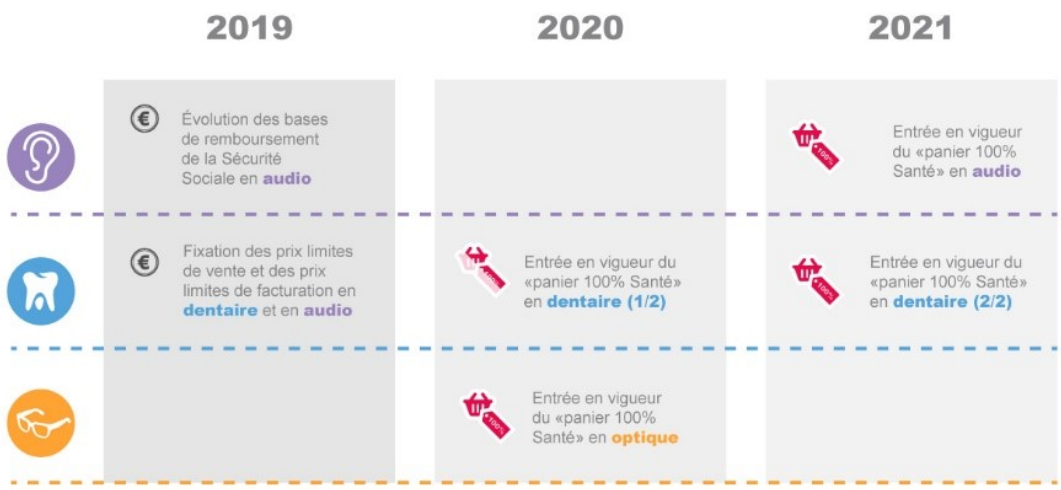

FIGURE 5.8 – 100 % santé illustration

Comparaison de contrat individuel et contrat collectif

	Contrat individuel	Contrat collectif
Bénéficaires	Salariés(sur complémentaire), étudiants,fonctionnaires,retraités, indépendants(TNS),auto entrepreneur	Salariés, adhérents d'association
Financement	Cotisation mensuelle en fonction du contrat	Cotisation mensuelle prise en charge à 50 % minimum
Choix	Libre de choisir l'assureur	Assureur choisi par l'entreperise
Garanties	Niveau de garantie en fonction du contrat	Panier de soins minimal en fonction du contrat choisi par l'entrepeise
Tarif	Les tarifs et les garanties dépendent de la formule choisie. Plus le montant est élevé, meilleur est le niveau de couverture.	Meilleur rapport qualité/prix(en général),négocié entre l'entreprise et la compagnie d'assurance pour l'ensemble des salariés.
Avantages	Libre choix de résillier (après un an d'adhésion)	Avantage fiscal pour l'entreprise

FIGURE 5.9 – Comparaison de contrat individuel et contrat collectif